دانشکده مهندسی

گروه مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد

روش جدید وزن دهی معنایی به کلمات در کاربردهای پردازش متن

تهیه و تنظیم:

حسین کامیار

استاد راهنما:

دکتر محسن کاهانی

استاد مشاور:

دکتر نادر جهانگیری

تابستان ۹۰

چکیده:

امروزه استفاده از وب به یکی از عناصر حیاتی زندگی انسانی تبدیل شده است. حتی در بسیاری از جوامع زندگی روزمره آدمی در صورت اختلال در وب دچار مشکلات اساسی می شود. به همین دلیل حجم اطلاعاتی متنی در سطح به طرز چشمگیری افزایش یافته است. حتی به صورت شهودی نیز می توان ادعا نمود که نرخ رشد اطلاعات متنی در دنیای امروزه از نرخ رشد داده به هر شکل دیگری مانند صوت، تصویر و ... بالاتر است. کاربران در میان این دریای داده های خام، همیشه به دنبال اطلاعات خاصی هستند. به این منظور احتیاج به پردازش متن و زبان که در حقیقت لایه بالایی متن می باشد، شدیداً وجود دارد. از این رو، در حال حاضر بیش از هر زمان دیگری نیاز به سیستم های پردازشگر زبان مانند، بازیابی کننده های اطلاعات، خلاصه سازها، مترجم ها و ... احساس می شود.

یکی از اولین گام ها در پردازش زبان وزن دهی به کلمات به عنوان ویژگی های قابل پردازش از یک متن می باشد. به همین دلیل تحقیقات فراوانی بر روی وزن دهی به کلمات به عنوان ابزارهای پیش خوشه بندی متون انجام می شود. هر چه دقت روش وزن دهی بالاتر باشد دسته بندی اولیه متون بهتر انجام شده و در نهایت دسته بندی اصلی آنها از دقت بهتری برخوردار خواهد بود. روش های مشهور فعلی وزن دهی به کلمات، معمولاً روش های آماری قرضی از دیگر کاربردهای خوشه بندی هستند، که مبتنی بر شمارش فرکانس کلمات می باشند. اما ویژگی های اصلی زبان، معنا و گرامر می باشد که توسط این روش ها قابل شمارش نمی باشند.

در این تحقیق یک روش جدید با رویکرد توجه به ویژگی های اصلی زبان برای وزن دهی به کلمات ارائه شده است. این روش با مبنا قرار دادن یک روش مشهور وزن دهی آماری به نام TF-IDF به تغییر پارامتر TF که یک پارامتر اندازه گیری فرکانس در سطح یک متن می باشد، می پردازد. این تغییرات از دو جنبه معنا توسط پایه قرار دادن یک تئوری زبانی به نام نظریه مرکزیت و گرامر با توجه به نقش گرامری کلمات در متن و توزیع آنها، انجام می گردند. همچنین جهت پر کردن خلأ توجه به تأثیر سراسری کلمات در مجموعه ای از متون در تغییر پارامتر TF به نقش گرامری کلمات در سطح کلیه متون نیز توجه شده است.

نتایج بدست آمده در پایان نامه به خوبی تأثیر روش پیشنهادی بر روش های پردازش زبان را نشان می دهد. یک چنین روشی تا میانگین ۱۱٪ نسبت به یک روش مشهور وزن دهی مانند TF-IDF بهبود در دقت در کاربردی مانند بازیابی اطلاعات رانشان می دهد.

کلمات کلیدی: وزن دهی به کلمات، بازیابی اطلاعات، TF-IDF، نظریه مرکزیت، نقش گرامری، انتقال معنایی، برچسب زنی گرامری، الگوریتم LSI، پردازش متن

فهرست مطالب

۱- مقدمه..... ۱۱

۱-۱- انگیزه.. ۱۱

۱-۲- راه حل پیشنهادی ۱۴

۱-۳- ساختار پایان نامه ۱۵

۲- مرور ادبیات.... ۱۸

۲-۱- بازیابی اطلاعات ۱۸

۲-۱-۱- سیستم بازیابی اطلاعات ۱۸

۲-۱-۲- مراحل بازیابی اطلاعات ۱۸

۲-۱-۳- معیارهای صحت و کیفیت در بازیابی ۲۰

۲-۲- الگوهای وزن دهی ۲۱

۲-۲-۱- وزن دهی با فرکانس کلمات ۲۱

۲-۳- تئوری بردارِکلمه و وزنِ کلمات کلیدی ۲۶

۲-۳-۱- مدل فضای برداری سالتون ۲۶

۲-۴- نظریه ی مرکزیت ۳۰

۲-۴-۱- شهود اصلی تئوری مرکزیت ۳۰

۲-۴-۲- تئوری مرکزیت و پارامترهای آن ۳۱

۲-۴-۳- اصطلاح شناسی و تعاریف ۳۲

۲-۴-۴- گذارها ۳۳

۲-۴-۵- ادعاهای اصلی ۳۴

۲-۴-۶- انتقال های معنایی نظریه مرکزیت ۳۴

۲-۵- پیش پردازش های زبانی ۳۵

۲-۵-۱- تشخیص زنجیره های مرجعیتی ۳۵

۲-۵-۲- برچسب زنی معنایی نقش کلمات ۳۶
۲-۵-۳- برچسب زنی نحوی لغات..... ۳۶
۲-۵-۴- حذف کلمات زائد ۳۷
۲-۵-۵- ریشه یابی کلمات ۳۷

۲-۶- کارهای مرتبط ۳۹

۲-۷- خلاصه فصل ۴۶

۳- روش پیشنهادی... ..... ۴۸

۳-۱- مقدمه..... ۴۸

۳-۲- ساختار سیستم پیشنهادی ۴۸

۳-۳- پیش پردازش های زبانی ۵۰
۳-۳-۱- پیش پردازش های مرتبط با الگوریتم پیشنهادی..... ۵۱

۳-۴- تشکیل ماتریس نهایی سند-لغت ۵۴
۳-۴-۱- تغییر اعمال شده در TF-IDF به لحاظ انتقال معنایی نظریه ی مرکزیت ۵۴
۳-۴-۲- حدس های شهودی درباره ی انتقال های معنایی نظریه ی مرکزیت..... ۵۴
۳-۴-۳- اثبات تجربی توانایی نظریه ی مرکزیت در تشخیص انسجام و برجستگی ۵۵
۳-۴-۴- ضرایب بدست آمده برای انتقال های معنایی نظریه ی مرکزیت برای کاربرد بازیابی اطلاعات. ..۵۵
۳-۴-۵- تعیین فرکانس مرتبط با هر کلمه در مجموعه ای از اسناد ۶۲
۳-۴-۶- محاسبه ی TF-IDF ۶۳

۳-۵- خلاصه ۶۴

۴- پیاده سازی و ارزیابی..... ۶۷

۴-۱- مقدمه ۶۷

۴-۲- پیاده سازی ۶۷
۴-۲-۱- تشخیص زنجیره های مرجعیتی ۶۷

۴-۲-۲- برچسب زنی معنایی نقش کلمات ۶۸
۴-۲-۳- برچسب زنی نحوی لغات ۷۰
۴-۲-۴- حذف کلمات زائد ۷۰
۴-۲-۵- ریشه یابی ۷۱
۴-۲-۶- تشخیص انتقال های معنایی میان جملات ۷۱

۴-۳- آزمایش درستی حدس های شهودی درباره انتقال های معنایی ۷۱
۴-۳-۱- خلاصه سازی بر مبنای نظریه ی مرکزیت ۷۱

۴-۴- ارزیابی روش وزن دهی پیشنهادی ۷۶
۴-۴-۱- توصیف داده ها ۷۶
۴-۴-۲- معیارهای وزن دهی مورد تحقیق ۷۷
۴-۴-۳- الگوریتم بازیابی مورد استفاده ۷۸
۴-۴-۴- آزمون ها ۷۹

۴-۵- خلاصه ۹۰

۵- نتیجه گیری و کارهای آتی..... ۹۳

۵-۱- نتیجه گیری ۹۳

۵-۲- کارهای آینده ۹۵

۶- منابع..... ۹۹

ضمائم ۱۰۴

۷- ضمائم. ۱۰۵

۷-۱- ضمیمه الف- مراحل پنج گانه یک سیستم بازیابی اطلاعات ۱۰۵
۷-۱-۱- ایندکس بندی ۱۰۵
۷-۱-۲- خطی سازی سند ها ۱۰۶
۷-۱-۳- فیلتر کردن ۱۰۷

۴-۱-۷- ریشه یابی ۱۰۸

۲-۷- ضمیمه ب- انواع مدل های نمایش مجموعه اسناد ۱۰۹

۱-۲-۷- مدل شمارِ کلمات ۱۰۹

۲-۲-۷- مدل فضای برداری کلاسیک ۱۱۱

۳-۲-۷- مدلهای برداری با فرکانسهای نرمال شده ۱۱۴

۴-۲-۷- مدل گلاسکو ۱۱۶

۳-۷- ضمیمه ج- اشکال نتایج آزمون های انتخاب ابعاد حقیقی ۱۱۶

۱-۳-۷- آزمون انتخاب ابعاد حقیقی، تابع PL ۱۱۶

۲-۳-۷- آزمون انتخاب ابعاد حقیقی، تابع AMDL ۱۱۷

۳-۳-۷- آزمون انتخاب ابعاد حقیقی، تابع PA ۱۱۹

فهرست شکل ها

شکل ۲-۱- یک سیستم بازیابی اطلاعات پایه ۱۸
شکل ۲-۲- *خطی سازی سند شامل A) حذف نشانه ها و فرمت ها B) توکن بندی C) فیلتر کردن D) ریشه یابی E) وزن گذاری* ۲۰
شکل ۲-۳- *شمای تصفیه شده از گذارهای مرکزیت* ۳۴
شکل ۳-۱- ساختار کلی سیستم وزن دهی پیشنهادی ۴۹
شکل ۴-۱- نمای گرافیکی از خروجی ابزار تشخیص زنجیره های مرجعیتی ۶۸
شکل ۴-۲- نمای گرافیکی از خروجی ابزار برچسب زنی معنایی نقش کلمات ۷۰
شکل ۴-۳- نمودار نتایج ارزیابی الگوریتم پیشنهادی با سیستم SweSum و ۱۱ سیستم موجود در DUC2001 [کام۸۹] ۷۶
شکل ۴-۴- خروجی SVD برای ماتریس کلمه-جمله ۷۹
شکل ۴-۵- نمودار مقادیر ابعاد حقیقی بازگردانده شده برای مجموعه داده MED به ازای معیارهای وزن دهی مختلف توسط تابع PL ۸۳
شکل ۴-۶- نمودارهای مقادیر ابعاد حقیقی بازگردانده شده برای مجموعه داده MED به ازای معیارهای وزن دهی مختلف توسط تابع AMDL ۸۴
شکل ۴-۷- نمودارهای افزایش کارایی الگوریتم LSI بر اساس معیارهای مختلف وزن دهی و توابع انتخاب ابعاد حقیقی متفاوت ۸۷
شکل ۴-۸- نمودار معیار MAP برای چهار مجموعه داده با هم به ازای معیار های متفاوت وزن دهی ۸۹
شکل ۷-۱- مراحل تبدیل یک متن به کلمات ایندکس شده ۱۰۶
شکل ۷-۲- نمودارهای مقادیر ابعاد حقیقی بازگردانده شده برای مجموعه داده های سه گانه به ازای معیارهای وزن دهی مختلف توسط تابع PL ۱۱۷
شکل ۷-۳- نمودارهای مقادیر ابعاد حقیقی بازگردانده شده برای مجموعه داده های سه گانه به ازای معیارهای وزن دهی مختلف توسط تابع AMDL ۱۱۹
شکل ۷-۴- نمودارهای نقطه برخورد ماتریس های تصادفی و واقعی مقادیر ویژه برای مجموعه داده های چهارگانه به ازای معیارهای وزن دهی مختلف توسط تابع PA ۱۲۱

فهرست جداول

جدول ۲-۱- تعدادی از فرمولهای متداولِ وزن محلی [SAL88].....۲۲
جدول ۲-۲- تعدادی از فرمول های متداول وزن سراسری [SAL88].....۲۳
جدول ۲-۳- تعدادی از فرمولهای متداول نرمال سازی [SAL88].....۲۴
جدول ۲-۴- الگوهای وزن دهی معمول [SAL88] ۲۶
جدول ۲-۵- مقایسه روش های معنایی وزن دهی به کلمات بر اساس ۷ ویژگی اصلی ۴۴
جدول ۳-۱- ضرایب تجربی بدست آمده برای انتقال های معنایی بر اساس میانگین رخداد آنها در مجموعه داده MED.....۵۷
جدول ۳-۲- ضرایب تجربی استفاده شده برای انتقال معنایی به دست آمده از آزمایش چندباره بر روی مجموعه داده MED.....۵۸
جدول ۳-۳- ضرایب استفاده شده برای نقش های گرامری، به دست آمده از آزمایش چندباره بر روی مجموعه داده MED.....۶۱
جدول ۴-۱- میانگین رخداد هر انتقال معنایی در DUC2001 [کام۸۹] ۷۳
جدول ۴-۲- مقدار تجربی ضریب β برای انتقال های معنایی [کام۸۹].....۷۴
جدول ۴-۳- اطلاعات کلی معرفی کننده چهار مجموعه داده ارزیابی.....۷۷
جدول ۴-۴- شش معیار وزن دهی مشهور به تفکیک پارامترهای وزن محلی و وزن سراسری آنها.....۷۷
جدول ۴-۵- وزن اختصاص داده شده به کلمه "Bacillus " توسط معیارهای مختلف در اسنادی که طی عمل بازیابی اطلاعات برای این کلمه بازگردانده می شوند.....۸۱
جدول ۴-۶- مقادیر ابعاد حقیقی برای چهار مجموعه داده توسط تابع PL به تفکیک معیار وزن دهی ۸۲
جدول ۴-۷- مقادیر ابعاد حقیقی برای چهار مجموعه داده توسط تابع AMDL به تفکیک معیار وزن دهی.۸۲
جدول ۴-۸- مقادیر ابعاد حقیقی برای چهار مجموعه داده توسط تابع PA به تفکیک معیار وزن دهی ۸۳

فصل اول:

مقدمه

۱– مقدمه

۱–۱– انگیزه

از آغاز به کار شبکه جهانی وب[1] افزایش بسیار زیادی در حجم اطلاعات موجود به فرمت الکترونیکی رخ داده است. در نتیجه، زمان لازم برای جستجوهای برخط به منظور دستیابی به اطلاعات خاص بسیار افزایش یافته است. این امر به نوبه خود، منجر به افزایش میزان ناکامی و خستگی در تعداد زیادی از کاربران شده است. جستجوگران غالباً در حجم زیادی از اطلاعات بازگشتی در رابطه با جستجو پیرامون یک موضوع خاص، غوطه ور می شوند. در حال حاضر مشکل اطلاعات اضافی تبدیل به یکی از اصلی ترین مشکلات در تعداد زیادی از وجوه زندگی امروزی شده است.

امروزه هم در فعالیتهای حرفه ای و هم در فعالیتهای آزاد و تفریحی، استفاده از وب به عنوان یک منبع ضروری برای دستیابی به اطلاعات و دانش در پهنه ی عظیمی از موضوعات، گسترش فراوانی یافته است. سهولت در اضافه کردن اطلاعات به صورت فرمت های غیر ساخت یافته و دیگر منابع دیجیتال به وب، یکی از دلایل افزایش زیاد اطلاعات در دسترس افراد می باشد. اسناد و متون با کمترین نیاز به دستکاری و تغییر، قابلیت قرارگیری و بارگذاری بر روی سایت ها را دارند. همچنین ساختار نسبتاً ساده ی صفحات وب و بی میلی نویسندگان در سازگاری خود با ساختارهای سخت و پیچیده برای تولید و ایجاد اسناد، باعث افزایش محبوبیت ایجاد و به اشتراک گذاری اسناد در سطح وب شده است. در مقابل، وجود حجم بالای اطلاعات، به ویژه اینکه درصد بالایی از آنها به فرمت های ساخت نیافته هستند، باعث ایجاد مشکلاتی در زمینه ی بازیابی اطلاعات در یک حوزه ی معین شده است.

سیستم های بازیابی اطلاعات[2] مرتبط است با اسناد زبان طبیعی، پرس و جو[3]ها و تلاش ها برای محدود کردن اطلاعات اضافی به وسیله ی بازگرداندن خودکار تنها اسنادی که مورد نیاز کاربر (پرس و جو) می باشند. انسان توانایی تفسیر درست عبارات و جملات مبهم در زبان طبیعی را دارد. در حالیکه، بازیابی خورکار اسناد زبان طبیعی دارای چندین مشکل می باشد. کلمات یکسانی که معانی مختلف دارند (تعدد معنایی[4]) و کلمات متفاوت با معانی یکسان (هم معنی[5]) دو پدیده در زبان طبیعی می باشند که جزء مشکلات بازیابی خودکار اطلاعات هستند. این پدیده ها می توانند باعث ایجاد ابهام در زبان طبیعی شوند، و بنابراین سیستم

---

[1] - World Wild Web
[2] - Information Retrieval System
[3] - Query
[4] - Polysemy
[5] - Synonym

های خودکار در حل این پیچیدگی ها دارای مشکلاتی هستند. زمانی که یک کلمه دارای چندین معنی و یا استفاده (تعدد معنایی) است، انتخاب معنی درستی که مورد نظر نویسنده بوده برای یک سیستم خودکار دشوار است. درمقابل، زمانی که چندین کلمه مرتبط با یک مفهوم (هم معنی) هستند، نگاشت کلمات متفاوت به مفهوم مناسب درست، مشکل می باشد. مطالب عنوان شده، جزو خصوصیات سیستم های بازیابی اطلاعات هستند که برای حل آنها روش های مختلفی و در عمق های مختلف وجود دارند. این روش ها خود می توانند باعث ایجاد مشکلاتی برای سیستم های بازیابی اطلاعات شوند [CUM08].

تعداد زیادی از سیستمهای بازیابی اطلاعات مبتنی بر مدل ساده ای از بردار فضای بازیابی می باشند [SAL75]. در این مدل کلمات پرس و جو (کلمات کلیدی) بر اساس میزان اهمیت آنها در احتمال حضور و تأثیرشان در اندازه گیری محتوای اطلاعات، وزن دهی می شوند. سپس این کلمات با لغات اسناد مختلف مطابقت داده می شوند. اسنادی که رخدادهای بیشتری از این کلمات وزن دهی شده در آنها موجود باشد، از امتیاز بالاتری برخوردار خواهند شد. امتیاز کلمات در هر سند برای برگرداندن امتیاز کلی آن سند، تجمیع می شود. در نهایت لیست امتیاز بندی شده ای از اسناد به کاربر برگردانده می شود، که سندی با بیشترین امتیاز در صدر این لیست قرار دارد. این روش همزمان با سادگی، دارای مزایایی مانند مؤثر بودن، کارایی و قابلیت پیاده سازی آسان می باشد. این نوع از روش ها اغلب به روش های "بسته ی کلمات"[1] مشهورند. این روش ها مبتنی بر این حقیقت هستند که، کلمات به صورت مستقل از یکدیگر رفتار می کنند.

روش های "بسته ی کلمات" از بعضی روشها برای نسبت دادن وزن به کلمات استفاده می کنند، که انعکاس دهنده ی میزان اهمیت کلمه در تشخیص موضوع سند می باشند. این روش های نسبت دهی وزن معولاً شماهای وزن دهی کلمات[2]و یا توابع وزن د هی کلمات[3]نامیده می شوند. در زمینه ی وب مدرن، گاهی اوقات این روش ها با نام توابع رتبه بندی[4]خوانده می شوند. در هر حال، وزن های داده شده به کلمات در بهبود انواع سیستم های بازیابی اطلاعات که از وزن دهی به کلمات بهره می برند، بسیار حیاتی می باشند [SAL88]. بنابراین این روش های وزن دهی به کلمات، پایه و مبنای بسیاری از موتورهای جستجوی کنونی می باشند. همچنین تئوری های پشت این روش ها برای درک مفاهیم نظری در بازیابی اطلاعات بسیار ضروری می باشند. بازیابی اطلاعات علاوه بر کاهش مشکل بازیابی زبان طبیعی به دنبال توسعه ی نظریه ی مهم بازیابی که ارتباط نامیده می شود، نیز می باشد. فهم طبیعت درست ارتباط به منظور توسعه ی سیستم هایی که به دقت این مفهوم را مدل می کنند، در صورتی که یک امر حیاتی محسوب نشود، جذاب می باشد.

---

[1] - Bag of words
[2] - Term weighting schemas
[3] - Term weighting functions
[4] - Ranking functions

اینکه ارتباط یک مفهوم یکتاست یا اینکه ارتباط یک مفهوم موازی در سایر زمینه های علمی است، یکی از سؤالات اساسی در بازیابی اطلاعات محسوب می شود [CUM08].

تعداد زیادی از منابع شامل شواهد بالقوه که می توانند برای استنتاج ارتباط یک پرس و جو داده شده از کاربر استفاده شوند، وجود دارد. معیارهای ضمنی یکی از این منابع ارزشمند است که کاربر نسبت به اینکه باید آنها را در اختیار سیستم قرار دهد، آگاه نیست. سیستم های بازیابی اطلاعات می توانند به صورت خودکار این دانش را جمع آوری کرده و بوسیله ی آن پرس و جوی کاربر را تکمیل کنند. اطلاعات زمانی و تاریخی می توانند در مورد عناوین خبری مرتبط مورد استفاده قرار گیرند و یا شواهد فضایی و جغرافیایی می توانند در هنگام جستجو به دنبال اطلاعات یک رویداد خاص در یک منطقه ی معین مورد استفاده قرار گیرند. به هر حال، مشکل اساسی تر و پایه ای تر در بازیابی اطلاعات بحث ابهام ذاتی زبان است. یک مرحله ی سودمند، آنهم در صورتی که واقعاً نیاز نباشد، توسعه ی سیستم هایی است که می توانند به صورت خودکار ابهام پرس و جو داده شده توسط کاربر را به منظور افزایش بهره وری، برطرف کنند. یک چنین روشی معمولاً تعدادی کلمه را که می توانند برای روال بازیابی مفید باشند، به پرس و جو داده شده اضافه می کند. این کلمات از دانش موجود در مورد زبانی که جستجو در آن صورت می گیرد، برگرفته می شوند. تکنیک های توسعه ی خودکار پرس و جو تلاش می کنند تا مشکل تفاوت های واژگانی (یا عدم تطابق) موجود در زبان را کاهش دهند. در تعدادی از این تکنیک های خودکار بسط پرس و جو، نمونه های زبان به منظور تشخیص شباهت های (هم معنی) کلمات مورد تحلیل قرار می گیرند. سپس می توان پرس و جو را بسط داده و اطلاعات مفید جدیدی به آن اضافه کرد.

روش ها و معیارهای وزن دهی به کلمات از اصلی ترین پایه های بسیاری از کاربردهای متن مانند پردازش زبان و یا بازیابی اطلاعات می باشند. مخصوصاً در کاربردهایی مانند بازیابی اطلاعات که احتیاج به خوشه بندی داده ها دارند، این روش ها از اهمیت بالاتری برخوردارند. چراکه وزن دهی به کلمات در حقیقت، مقداردهی به ویژگیهای مجموعه داده ی متنی که کلمات هستند، می باشد. از این رو یک وزن دهی خوب به کلمات باعث می شود، تا یک خوشه بندی ابتدایی بر روی مجموعه ی داده های متنی بوجود آید. اگر این خوشه بندی ابتدایی باعث شود که ابهام بالای مجموعه ی داده های متنی (که خصوصاً به دلیل دو چالش عمده تعدد معنایی و هم معنی به وجود آمده است) که این مجموعه را به مجموعه ای غیر خطی و فوق العاده ناهمگون تبدیل کرده است، تا حدی برطرف شود، روش های وزن دهی بالاترین میزان تأثیر را در کاربردهای پردازش متن خواهند داشت.

در حال حاضر روش های خوشه بندی در مورد مجموعه داده های متنی از کارایی خوبی برخوردار نیستند [LUO11]. البته روش های متعددی که بر روی داده های غیر خطی کار می کنند، وجود دارد. اما روش

های کارایی که بر روی مجموعه داده های فوق العاده ناهمگون ولی دارای ارتباطات ذاتی – مانند معنا در متن- نتایج خوبی بدهند، وجود ندارد.

روش های وزن دهی به کلمات، معیارهای شمارش آماری کلمات – که از کاربردهای دیگر خوشه بندی که بر روی مجموعه داده هایی مانند تصویر، صوت و ... اقتباس شده اند- و یا وزن دهی به خواص ذاتی کلمات مانند معنا، نقش گرامری و ... هستند.

در حال حاضر قریب به اتفاق روش های موجود در وزن دهی به کلمات، مبتنی بر ویژگی های آماری مجموعه ی داده ها – که در اینجا داده های متنی می باشند – قرار دارند. یکی از دلایل عمده ی این امر، قرضی بودن این روش ها می باشد. روش های وزن دهی به کلمات که معمولاً مبتنی بر قواعد آماری می باشند، از کاربردهای مختلف داده کاوی در حوزه های دیگر اطلاعاتی مانند تصویر، صوت و ... به قرض گرفته شده اند. در این حوزه های اطلاعاتی، تقریباً تمامی خواص ویژگی ها، توسط شاخصه های آماری قابل تعریف می باشند. اما در حوزه ی متن خواص اصلی مانند هم معنی ها و کلمات دارای تعدد معنایی به صورت کامل قابل تعریف توسط شاخصه های آماری نمی باشند [KAN05]. به دلیل این ویژگی ذاتی حوزه ی اطلاعات متنی و همچنین قرضی بودن اکثر قریب به اتفاق روش های آماری وزن دهی به کلمات، توجه به ویژگی های زبانی و معنایی، از محبوبیت بالایی برخوردار شده است. یک دلیل مهم در ناامیدی نسبت به روش های کنونی، عدم کارایی بالای آنها در معیارهای کارایی در کاربردهای مانند بازیابی اطلاعات می باشد [KAN05].

توجه مطلق به ویژگی های آماری و عدم توجه کافی نسبت به ویژگی های ذاتی زبان باعث شده است، که معیارهای کنونی از اقبال خوبی در خوشه بندی اولیه داده ها و رفع ناهمگونی فوق العاده بالای این داده ها، برخوردار نباشند [LUO11].

۲–۱– راه حل پیشنهادی

می توان خواص ذاتی زبان را به سه دسته ی عمده ی معنایی، گرامری/نحوی و ابر زبانی تقسیم کرد. هر کلمه در یک متن شامل دو ویژگی قابلیت حمل معنا و داشتن جایگاه گرامری در یک جمله می باشد. خاصیت های ذاتی ابر زبانی مانند تغییر لحن در هنگام صحبت نیز جزء خواصی از زبان هستند، که به مقدار زیاد در تشکیل ویژگی های یک قطعه ی گفتار، نقش ایفا می کنند. اما این خواص قابل پردازش و یا به سختی قابل پردازش و تفسیر در قالب متن هستند. از این رو امروزه دو خاصیت معنا و نقش های گرامری/نحوی در وزن دهی به کلمات مورد توجه گسترده هستند.

به همین منظور در تحقیق پیش رو سعی شده است، تا با انتخاب یکی از بهترین روش های آماری شمارش فرکانس و وزن دهی به کلمات و تغییر معیار شمارش فرکانس بر حسب ویژگی های معنایی و زبانی تا حد زیادی خلأ موجود پر شود.

در این پایان نامه برای پرداختن به ویژگی های معنایی، از تئوری زبانی به نام نظریه مرکزیت[1] که به دنبال درک، تفسیر و شناسایی نحوه ی گردش معنا در داخل یک قطعه ی گفتار است، استفاده خواهیم کرد. البته این تئوری قابلیت درک و تفسیر گردش معنایی در مجموعه ی زیادی از داده های را ندارد. اما برای درک معنا در مجموعه ی بزرگی از اسناد نیز تنها می توان به استفاده از مجموعه هایی از واژگان به نام هستان شناسی[2] واژگان دل بست. این امر مستلزم سالها صرف وقت دانش بشری برای ساخت این هستان شناسی ها می باشد و به عنوان یک روش هوشمند کامل، چندان مورد تأیید قرار نمی گیرد. اما نظریه مرکزیت، مدعی تشخیص درستی از معنا به صورت محلی، بدون احتیاج به دانش قبلی می باشد. این نظریه، برای نیل به این هدف، دست به تعریف پارامترهایی برای جملات یک متن می زند. این پارامترها بر اساس نقش گرامری کلمات و همچنین کلمات یکسان تکرار شده در جملات متوالی، برای هر جمله تعریف می شوند. در حقیقت، تشکیل این پارامترها، مبنای قابلیت استفاده از نظریه مرکزیت در الگوریتم ها می باشند. سپس در این نظریه بر اساس پارامترهای تعریف شده، انتقال های معنایی میان جملات متوالی تعریف می گردند. در تحقیق پیش رو تنها به چهار عدد از این انتقال ها با نام های Continue، Smooth-shift، Retain و Rough-shift بسنده شده است. در این پایان نامه بر اساس انتقال انجام شده میان دو جمله، وزن جدیدی به کلمات آن جمله اختصاص داده می شود.

سپس برای پرداختن به ویژگی گرامری/نحوی کلمات در زبان، به تأثیر نقش گرامری کلمات در یک متن می پردازیم. برای نقش های مختلف گرامری اولویتی ارائه می شود، که بر اساس این اولویت و نیز نحوه توزیع آن نقش گرامری در سطح یک متن، به کلمات مختلف وزن تخصیص داده می شود.

یک خلأ قابل اشاره تا این مرحله، عدم توجه به تأثیر سراسری کلمات می باشد، که به خوبی در روش های آماری وزن دهی به کلمات مورد توجه قرار می گیرد. از این رو، در این تحقیق برای کلمات با توجه به نقش گرامری آنها و نحوه توزیع این نقش گرامری در کل مجموعه ی اسناد نیز وزن جدیدی را اختصاص می دهیم.

۳-۱- ساختار پایان نامه

[1] - Centering theory
[2] - Ontology

در این پایان‌نامه براساس نظریه مرکزیت و توجه به نقش گرامری کلمات، روشی جدید برای وزن دهی به کلمات ارائه می‌شود. ادامه‌ی این نوشتار شامل ۴ فصل می‌باشد که عبارتند از:

فصل اول: در این فصل در ابتدا مروری به بررسی بازیابی اطلاعات به عنوان کاربرد مورد هدف از وزن دهی به کلمات در این پایان نامه خواهیم پرداخت. در بخش دوم به بررسی مفهوم وزن دهی به کلمات و عناصر تشکیل دهنده آن پرداخته شده است. در بخش سوم این فصل یک مطالعه ی اجمالی از نظریه مرکزیت را برای آشنایی بیشتر با این تئوری و پایه های تئوریک آن مشاهده خواهد شد. در بخش چهارم مروری بر کارهای انجام شده بر روری روش های وزن دهی به کلمات چه از دیدگاه آماری، چه از دیدگاه معنایی و چه از دیدگاه گرامری/نحوی انجام خواهد شد.

فصل دوم: در این فصل ابتدا به بحث در مورد پیش پردازش های زبانی مورد نظر و ساخت ماتریس ابتدایی برای هر متن و کل مجموعه ی اسناد پرداخته خواهد شد. همچنین شمای کلی از روش پیشنهادی ارائه خواهد شد. سپس با انتخاب معیار وزن دهی TF-IDF راه حل های ارائه شده برای تغییر پارامتر TF از دو دیدگاه معنایی – نظریه مرکزیت- و گرامری مورد بحث و بررسی قرار می گیرند. در این بخش همچنین یک تجربه آزمایشی برای اثبات شهودهای ارائه شده در مورد نظریه مرکزیت مورد بحث قرار گرفت.

فصل سوم: در این بخش ابتدا به نحوه ی پیاده سازی و یا معرفی ابزارهای لازم برای گام های پیش پردازشی متن می پردازیم. سپس به معرفی اجمالی داده های مورد آزمایش خواهیم پرداخت. در زیر بخش های بعدی به معرفی سه آزمون رایج در ارزیابی روشهای وزن دهی به کلمات و ارائه ی نتایج به دست آمده از مقایسه ی روش پیشنهادی با شش روش وزن دهی به کلمات مشهور دیگر بر اساس این سه آزمون خواهیم پرداخت. افزایش کارایی بر اثر استفاده از معیار پیشنهادی در آزمون های سه گانه مشهود می باشد.

فصل چهارم: در این فصل به نتیجه‌گیری روشهای ارائه شده خواهیم پرداخت و پیشنهادهایی برای کارهای آتی ارائه خواهیم داد.

فصل دوم:

مرور ادبیات

۲- مرور ادبیات

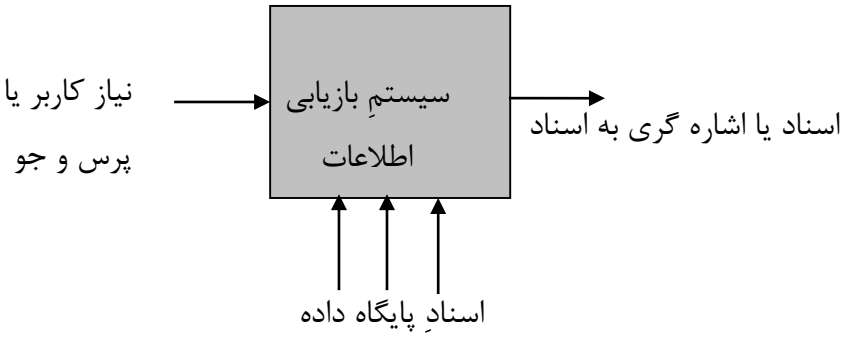
۲-۱- بازیابی اطلاعات

۲-۱-۱- سیستم بازیابی اطلاعات

سیستم بازیابی اطلاعات[1]در مجموعه سندها[2]یی که به فرمتهای گوناگون در پایگاه داده وجود دارند(حال این پایگاه داده ممکن است پایگاه دادهٔ رابطه ای مستقل[3]یا پایگاه دادهٔ شبکه ای ابر متن[4]مثل اینترنت باشد) [93], [DAT81]، به دنبال اطلاعات مطلوبِ کاربر می گردد. جستجوی اطلاعات به دو صورت امکان پذیر است[ROB76]:

- نگاه کردن کلی[5]در مواردی است که خواستهٔ کاربر دقیق نباشد یا اینکه علایق کاربر گسترده باشد.
- جست و جو کردن[6]در مواردی است که خواستهٔ کاربر دقیق باشد.

در شکل ۲-۱ شمایی از یک سیستم بازیابی اطلاعات پایه آمده است:



شکل ۲-۱- یک سیستم بازیابی اطلاعات پایه

۲-۱-۲- مراحل بازیابی اطلاعات

در بازیابی اطلاعات چندین مرحلهٔ مهم وجود دارد [COH95]:

۱. ایندکس بندی[7]

---

[1]- Information Retrieval(IR) System
[2]- Document
[3]- Relational Stand-Alone Database
[4]- Hypertext Networked Database
[5] - Browsing
[6] - Quering
[7] - Indexing

۲. خطی سازی سند[۱]
   - حذف نشانه ها و فرمت[۲]
   - نشان گذاری[۳]
۳. فیلتر کردن[۴]
۴. ریشه یابی[۵]
۵. وزن دادن[۶]
۶. رتبه بندی[۷]

۱-۲-۱-۲- وزن دهی

از بین مراحل شش گانه بالا و با توجه به رویکرد این پایان نامه، تنها به تشریح گام وزن دهی خواهیم پرداخت. توضیحات بیشتر جهت آشنایی با مراحل پنج گانه دیگر در ضمیمه الف ارائه شده است.

وزن دهی مرحلهٔ نهائی در اکثر برنامه های ایندکس بندی برای بازیابی اطلاعات است. کلمات براساس مدلی وزن دهی می شوند که این مدل ممکن است بر اساس وزن دهی محلی و یا سراسری و یا مخلوطی از هر دو باشد [ROB76].

اگر وزن دهی محلی[۸]مورد استفاده باشد، آنگاه وزن کلمه ف تعداد تکرار (فرکانس) کلمه می باشد (tf). حال اگر وزن سراسری[۹]مورد استفاده باشد وزن کلمه با مقدارِIDF بیان می گردد. اصلی ترین و اساسی ترین الگوی وزن دهی مدلی است که در آن از هر دوی وزن محلی و وزن سراسری استفاده می شود (وزن کلمه= tf * IDF) که به طور معمول از آن به عنوان وزن دهیtf*IDF یاد می شود [ROB76]. در ادامه به مدلهای وزن دهی و تشریح آنها نیز می پردازیم. درشکل۲-۲ سریِ مراحل ایندکس بندی با مثال مشخص شده است.

---
[1] - Document Linearization
[2] - Markup and Format Removal
[3] - Tokenization
[4] - Filtration
[5] - Stemming
[6] - Weighting
[7] - Ranking
[8]- Local Weighting
[9]- Global Weightnig

Interactive
modifies q
from a use
expansion
automatic

Mar

interactive
modifies q
automatic
expands q

(StopWord

automat 2

expand 1

modify 1

term 1

شکل ۲-۲- *خطی سازی سند شامل A) حذف نشانه ها و فرمت ها B) توکن بندی C) فیلتر کردن D) ریشه یابی E) وزن گذاری*

۳-۱-۲- معیارهای صحت و کیفیت در بازیابی

قبل از آنکه به بررسی مدلهای وزن دهی و تئوری بردار کلمه بپردازیم دو معیار Precision و Recall که جزو معیارهای ارزیابی صحت و کارایی سیستم های بازیابی اطلاعات به شمار می روند، می پردازیم:

Precision: نسبت سندهای بازیابی شدهٔ مرتبط به تمامی سندهای بازیابی شده می باشد [ROC66]:

$$\text{Precision} = \frac{\left|\{\text{Re}\,levantDocuments\} \cap \{\text{Re}\,trievedDocuments\}\right|}{\left|\{retrievedDocuments\}\right|} \quad (۲-۱)$$

Recall: نسبت سندهای بازیابی شدهٔ مرتبط به تمامی سندهای مرتبط می باشد [ROC66]:

$$\text{Recall} = \frac{\left|\{\text{Re}levantDocuments\} \cap \{\text{Re}trievedDocuments\}\right|}{\left|\{\text{Re}levantDocuments\}\right|} \quad (۲-۲)$$

F-measure: میانگین همساز[1]وزن دهی شده از precision وrecall می باشد [ROC66]:

$$\text{F} = 2 * \frac{\text{precision} * \text{recall}}{(\text{precision} + \text{racall})} \quad (۲-۳)$$

که F1 نیز نامیده می شود چونکه precision و recall به یکسان وزن دهی شده اند.

فرمول کلی برای مقدار حقیقیِ نامنفیِ α برابر است با [ROC66]:

$$\text{F}_{\alpha} = (1+\alpha) * \frac{\text{precision} * \text{recall}}{(\alpha * \text{precision} + \text{racall})} \quad (۲-۴)$$

Mean Average Precision: میانهٔ میانگینِ دقتها می باشد پس از آنکه سندهای مرتبط بازیابی شد.(بر روی مجموعه ای از کوئریها) [ROC66].

$$\text{Ave}P = \frac{\sum_{r=1}^{N} \left(P(r) * rel(r)\right)}{\text{number of relevant Documents}} \quad (۲-۵)$$

بطوریکه r رتبه، N تعداد سند بازیابی شده، $rel(r)$ تابع باینری بر روی ارتباطِ رتبهٔ داده شده و $P()$ دقت در cut-off rankِ داده شده می باشند.

۲-۲- الگوهای وزن دهی

هدف از الگوهای وزن دهی به کلمات، طبقه بندی کلمات ایندکس بندی شده بوسیلهٔ وزن دادن به آنها بر اساس تأثیر آنها در بهبود Recall و precision در امر بازیابی است.

در وزن دهی به کلمات مدل های مختلفی وجود دارد، که در بازیابی اطلاعات سنتی، شاخص ترین آنها مدل های بولی، احتمالی و فضای برداری می باشند. در این تحقیق، تمرکز بر روی مدل فضای برداری و تغییر یکی از معروفترین معیارهای موجود در این مدل، یعنی TF-IDF می باشد.

۱-۲-۲- وزن دهی با فرکانس کلمات

[1]- Harmonic Mean

الگوهای وزن دهی بر مبنای فرکانس کلمات، خواص تکرار کلمات در مجموعهٔ اسناد را معین می کنند [SAL88]. وزن نهایی کلمات با سه فاکتور بیان می گردد:

- L_{ij}: وزن محلی کلمهٔ i در سند j.
- G_i: وزن سراسری کلمهٔ i.
- N_j: فاکتور نرمال سازی برای سند j.

وزن کلمهٔ i در سند j توسط عبارت زیر بیان می گردد [CHI99]:

$$d_{ij} = L_{ij} G_i N_j \quad (۲-۶)$$

وزن های محلی تابعی از تکرار هر کلمه در یک سند است. قانون آن است که کلمه ای که در یک سند بیشتر تکرار شود، آن کلمه برای آن سند مهمتر است.

درجدول ۲-۱ تعدادی از فرمولهای متداولِ وزن محلی آمده است.

جدول ۲-۱- تعدادی از فرمولهای متداولِ وزن محلی [SAL88]

علامت مخفف	نام	فرمول
BNRY	**Binary**	**1** if $f_{ij} > 0$ **0** if $f_{ij} = 0$
FREQ	**Within-document frequency**	F_{ij}
LOGA	**Log**	**1+ log** f_{ij} if $f_{ij} > 0$ **0** if $f_{ij} = 0$
LOGN	**Normalized log**	**(1+log** f_{ij}**) / (1+log** a_j**)** if $f_{ij} > 0$ **0** if $f_{ij} = 0$
ATF1	**Augmented normalized term frequency**	**0.5 + 0.5 (** f_{ij} / x_i**)** if $f_{ij} > 0$ **0** if $f_{ij} = 0$

به طوریکه f_{ij} فرکانس کلمهٔ i در سند j، a_j فرکانس متوسط کلماتی که در سند j وجود دارند و x_j فرکانس ماکزیمم در بین کلمات موجود در سند j می باشند.

ساده ترینِ فرمولها BNRY وFREQ می باشد. با این وجود، BNRY بین کلماتی که متناوباً در یک سند تکرار می شوند با آنهایی که فقط یکبار تکرار شده اند، فرقی نمی گذارد. در حالیکه FREQ به کلماتی که متناوباً در

سند تکرار شده اند اگرچه برای سند مهم نباشند، وزن بیشتری می دهد.به منظور کاهشِ فرق گـذاریِ FREQ بین کلمات با تناوب بالا و کلمات کمتر تکرار شده دو راه حل وجود دارد [SAL88]:

۱. استفاده از لگاریتم: LOGA و LOGN. LOGN نسخهٔ نرمـال شـده از LOGN اسـت. فـاکتور نرمـال سازی $1+\log a_j$ است که وزن کلمهٔ(موجود یا ناموجود) با فرکانس متوسط در سند j را نشـان مـی دهد. LOGN برای آندسته از الگوهای وزن دهی استفاده می شود که از فاکتور سـوم اسـتفاده نمـی کند(فاکتور نرمال سازی سند).
۲. ATF1: به کلماتی که در سند ظاهر شده اند، وزن استانداردِ ۰,۵ داده می شود. سـپس بـه کلمـاتی که تناوب تکرار بالا در سند دارند، وزنی اضافی داده می شود.

وزن های سراسری تابعی از تعداد تکرار هر کلمه در کل مجموعهٔ اسناد است. قانون آن است کـه کلمـاتی کـه تنها در عدهٔ کمی از سندها ظاهر می شوند تمایزدهنده های بهتری نسبت به کلماتی هسـتند کـه در تعـدادی زیادی از اسناد ظاهر شده اند.

جدول۲-۲ تعدادی از فرمول های متداول وزن سراسری را نشان می دهد.

جدول ۲-۲- تعدادی از فرمول های متداول وزن سراسری [SAL88]

علامت مخفف	نام	فرمول
IDFB	**Inverse document frequency**	**log [(N / ni]**
IDFP	**Probabilistic inverse**	**log [(N – ni) / ni]**
ENPY	**Entropy**	$1+\sum_{j=1}^{N}\frac{\frac{f_{ij}}{F_i}\log\frac{f_{ij}}{F_i}}{\log N}$
IGFF	**Global frequency IDF**	**Fi / ni**
NONE	**No global weight**	**1**

به طوریکه N تعداد اسناد در کل اسناد، n_i تعداد اسنادی که کلمهٔ i در آنها ظاهر شده است و F_i فرکانس تکرار کلمهٔ i در سرتاسر مجموعه اسناد= $\sum_{j=1}^{N} f_{ij}$ می باشند.

هر کلمهٔ i در یک سند با احتمال p ممکن است ظاهر گردد. از آنجا که احتمال با فرکانس نسبی محاسبه مـی گردد بنابراین داریم p=ni/N.

با توجه به اینکه کلمهٔ با تعداد تکرار بالا تمایز دهندهٔ خوبی نیست، IDFB وزن کمی توسط فرمول زیر به کلمات با احتمال حضور بالا می دهد [SAL88]:

$$\log(1/p) = \log(N/n_i) \quad (۲-۷)$$

IDFP ارزیابی مشابهی با محاسبهٔ زیر انجام می دهد [SAL88].

$$\log[(1-p)/p] = \log[(N-n_i)/n_i] \quad (۲-۸)$$

تفاوت این دو آنجاست که IDFP به کلماتی که در بیش از نیمی از اسناد مجموعه ظاهر شده اند وزنی منفی می دهد((1-p)/p < 1) در حالیکه کمترین وزنی که یک کلمه در IDFB ممکن است داشته باشد صفر است.

ENPY انتروپی اطلاعات هر کلمه را محاسبه می کند. اگر یک کلمه در هر سند یکبار ظاهر شود، وزنِ $1+\sum_{j=1}^{N}\frac{-1}{N}=0$ به خود می گیرد..هر ترکیب دیگری از فرکانسهای کلمه در اسناد، وزنی بین صفر و یک در ENPY می گیرد.

IGFF به کلماتی که در تعداد کمی از اسناد متناوباً ظاهر شده اند، وزن های بزرگی می دهد. کمترین مقدار ممکن برای وزن درIGFF، یک است که به کلماتی که فرکانس تکرار آنها در اسنادی که ظاهر شده اند فقط یکبار است داده می شود [SAL88].

فاکتور نرمال سازی برای تفاوتها در اندازهٔ اسناد در نظر گرفته می شود. در اینجا قانون این است که کلمه ای که در یک سند کوچک و یک سند بزرگ به یک اندازه تکرار شده است، احتمالاً برای سند کوچکتر مهم تر است [SAL88].

نرمال سازی بردارهای اسناد تضمین می کند که اسناد، مستقل از بزرگی و اندازهٔ شان بازیابی می شوند.

جدول۲-۳ تعدادی از فرمولهای متداول نرمال سازی را نشان می دهد.

جدول ۲-۳- تعدادی از فرمولهای متداول نرمال سازی [SAL88]

علامت مخفف	نام	فرمول
DL	**Document length**	$\frac{1}{l_j}$
COSN	**Cosine normalization**	$\frac{1}{\sqrt{\sum_{i=0}^{l_j}(G_i L_{ij})^2}}$

$\frac{1}{(1-slope)\,pivot + slopel_j}$	**Pivoted unique normalization**	PUQN

به طوریکه l_j تعداد کلمه های یکتا در سند j و $slope$ = ۰,۲(این مقدار بصورت تجربی مشخص شده است).

برای هر سند، COSN بر بزرگی بردارِ سندِ وزن دهی شده تقسیم می گردد. بنابراین، بزرگی بردارهای سند وزن دهی شده در نهایت

$$\sum_{j=1}^{l_j}\left(\frac{G_i L_{ij}}{\sqrt{\sum_{j=1}^{l_j}\left(G_i L_{ij}\right)^2}}\right)^2 = 1$$

خواهد بود.

COSN در اسناد بزرگتر، به کلمات یکتا وزن کلمهٔ کوچکتری می دهد بنابذاین در بازیابی، اسناد کوچکتر نسبت به اسناد بزرگتر برتری دارند.

PUQN سعی در رفع این مشکل دارد. قاعدهٔ اصلی که توسط این روش استفاده می شود؛ سعی در کاهش دادن تفاوت بین احتمال آنکه یک سند مربوط باشد و احتمال آنکه یک سند بازیابی شود، می باشد. این تفاوت ممکن است بخاطر اندازهٔ سند ایجاد شده باشد و PUQN سعی در تصحیح این عامل دارد [SAL88]:

- در ابتدا فاکتور نرمال سازیِ متفاوتی استفاده می شود.(مثلاً 1/Lj)
- مجموعه ای از اسناد بازیابی شوند.
- منحنی های بازیابی وارتباط در مخالفت با طول سند کشیده می شوند.
- نقطه ای که دو منحنی با هم برخورد می کنند نقطهٔ لولایی[1] نامیده می شود.
- اسنادی که در سمت چپ نقطهٔ لولایی قرار دارند، احتمال آنکه مرتبط باشند از احتمال آنکه بازیابی شوند بیشتر است.
- حال l_j با $(1 - slope)\,pivot + slope\,l_j$ جایگزین می شود.

هرالگوی وزن دهی، فرمولهای متفاوتی را برای محاسبهٔ این سه فاکتور با هم ترکیب می کند. به طور نوعی ممکن است فرمولهای مورد استفاده برای وزن دهیِ بردارهای کوئری با فرمولهای مورد استفاده برای بردارهای سند فرق داشته باشند. جدول۲-۴ الگوهایی که بیشتر مورد استفاده است را نشان می دهد.

[1] - pivot

جدول ۲-۴- الگوهای وزن دهی معمول [SAL88]

نام الگو	وزن کوئری	وزن سند
Log – entropy	LOGA ENPY	LOGA ENPY COSN
IGFF – entropy	ATF1 ENPY	LOGA IGFF COSN
Raw term frequency	FREQ NONE	FREQ NONE NONE
Raw cosine	FREQ NONE	FREQ NONE COSN
SMART "Itc"	LOGA IDFB	LOGA NONE COSN
Variation of SMART	LOGA IDFP	LOGA NONE COSN
Best fully weighted	ATF1 IDFB	FREQ IDFB COSN
Best probabilistic	BNRY IDFP	ATF1 NONE NONE
Pivoted unique new norm weight	LOGA IDFB	LOGN NONE PUQN

۲-۳- تئوری بردارِ کلمه و وزنِ کلمات کلیدی

۲-۳-۱- مدل فضای برداری سالتون

مدل فضای برداری سالتون[1] مشهور ترین مدل کنونی برای نمایش بردار اسناد و متون جهت پردازش می باشد. به همین جهت در این تحقیق، به توضیحاتی اجمالی درباره آن خواهیم پرداخت و از این مدل برای نمایش بردارهای اسناد استفاده خواهیم کرد. چند مدل معروف دیگر در این زمینه، جهت آشنایی بیشتر در ضمیمه ب ارائه شده اند.

سیستمهای بازیابی اطلاعات با در نظر گرفتن موارد زیر به کلمات وزن می دهند [SAL88]:

۱. اطلاعات محلی از سندهای مجزا.
۲. اطلاعات سراسری از مجموعهٔ سندها.

به علاوه سیستمهایی که به لینک ها وزن می دهند، به منظور درک روشنِ درجهٔ اتصال بین سندها از اطلاعات گرافِ وب[2] استفاده می کنند.

در مطالعات بازیابی اطلاعات، مدلِ وزن دهی کلاسیک، همان مدل فضای برداری سالتون است که به طور معمول به "مدل بردار کلمه"[3] مشهور است. وزن دهی در این مدل از رابطهٔ زیر محاسبه می شود [SAL88]:

[1] - Salton's Vector Space Model
[2] - Web Graph
[3] - Term Vector Model

$$\text{وزن کلمه} = w_i = tf_i * \log\left(\frac{D}{df_i}\right) \quad (۲-۹)$$

که در آن:

- tf_i = فرکانس(تعدادِ کلمه) یا تعداد باری که کلمهٔ i در یک سند ظاهر شده است.
- df_i = فرکانسِ سند یا تعداد سندهایی که کلمهٔ i در آنها ظاهر شده است.
- D = تعداد سندهای موجود در پایگاه داده

برخی مدلها که بردار کلمات را از اسناد و کوئری ها استخراج می کنند از تساوی۲-۹ بدست آمده اند.

۱-۱-۳-۲- وزن های محلی

در تساویِ۲-۹، w_i با افزایش tf_i افزایش می یابد که این امر باعث خطاپذیری این مدل با سوءِاستفاده از تکرار کلمات است. یک نمونه از چنین سوءِ استفاده ای مشکلی به نام هرزنامه کلمات کلیدی[1]در اینترنت است. هرزنامه روشی است که برخی از صفحات برای احراز رتبهٔ بالاتر در موتورهای جستجو پیش میگیرند و آن بدین صورت است که با تکرار بیش از حد کلمات، بطور عمومی سعی در برهم زدن تعادل و در نتیجه فریب موتورهای جستجو را دارند [SAL88]. بنابراین

۱. در بین سندهای با طول یکسان، آنهایی که نمونه های بیشتری ازکلمهٔ پرس و جو شده را دارند در فرآیند بازیابی مطلوبترند.
۲. در سندهای با طول متفاوت، از آنجائیکه سندهای با طول بیشتر احتمال بیشتری برای داشتن نمونه های بیشتر از کلمه پرس و جو شده را دارند، بنابراین سندهای با طول بیشتر مطلوبترند.

۲-۱-۳-۲- وزنهای سراسری

در تساویِ۲-۹ عبارتِ $\log\left(\frac{D}{df_i}\right)$ به عنوان فرکانسِ سند معکوس[2]یا IDF_i شناخته می شود. این در واقع اندازهٔ حجم اطلاعاتِ(و انتروپی) مرتبط با کلمه در مجموعهٔ اسناد است. در طول سالیان تغییرات زیادی در تساویِ۲-۹ اعمال شده است.در اصطلاح، عبارتِ "مدلtf*idf" به معنای مدلی است که بر مبنای تساویِ۲-۹ بنا شده است [SAL88].

---

[1] - Keyword Spamming
[2] - Inverse Document Frequency

تساویِ۲-۹ نشان می دهد که w_i با افزایش df_i کاهش می یابد. برای مثال اگـر در یـک پایگـاه داده بـا ۱۰۰۰ سندتنها ۱۰ سندشامل کلمهٔ "ایران" باشند آنگاهIDF برای این کلمه از عبارت ۲-۹ به دست می آید.

$$\mathrm{IDF} = \log\left(\frac{1000}{10}\right) = 2$$

حال اگر تنها یک سند شامل این کلمه باشد، آنگاه $\mathrm{IDF} = \log\left(\frac{1000}{1}\right) = 3$.

بنابراین کلماتی که در سندهای متعددی ظاهر می شود(مثل کلمات با استعمال زیاد و کلمات ایست و حـروف ربط) وزن کوچکی پیدا می کنند در حالیکه کلمات غیر مشترک که تنها در عدّهٔ معدودی از سندها مـی آینـد، وزن بیشتری می گیرند و به این خـاطر اسـت کـه کلمـات متـداول(مثـلِ "بـا"، "از"، " بـه" ,"to", "For" ,"from") برای جداسازی سندهای مرتبط از سندهای غیر مرتبط چنـدان مفیـد نیسـتند. دو سـری در کـار بازیابی مفید نیستند؛ کلماتی که وزن قابل قبول دارند آنهایی نیستند که خیلـی متـداول و یـا خیلـی کمیـاب باشند، به عبارتی دیگر بردار کلماتِ آنها بسیار نزدیک به یا بسیار دور از بردارِ کوئری نیستند.

ملاحظه. در نمایش فضای برداری، هنگامی که کلمات غیر مشترک در اسناد و کوئری ها یافت می شوند بـردارِ کلمات متناظر(بردار کوئری و سند) بسیار به هم نزدیکند.

بعد از نمره دهی و مرتب سازی نتایج، گرایش سیستم به آن است که به این اسناد رتبهٔ بالایی بدهد، در حـالی که نتایج جستجوی کمی بازگردانده می شود. این گرایش نشان دهنده آن است که، رتبه بنـدیِ مطلـق نتـایج که از این بردار کلمات ناشی می شود، همیشه وجه مشخصهٔ خوبی برای ارتباط نیست؛ به زبان ساده 10 بودن در بین 5'000'000 از نتایج همانند 1 بودن در بین 5نتیجه نیست.

۳-۱-۳-۲- چگالی کلمات کلیدی

از تساوی۲-۹ آشکار است که وزن کلمات کلیدی[1]تحت عوامل زیر تحت تأثیر است [FAL95]:

۱. *شمار*[2]کلمات کلیدی
۲. حجم اسناد منتخب از پایگاه داده.

بنابراین این تصور عمومی که وزن کلمات توسط" میزان چگالیِ کلمات کلیدی" می تواند قابل محاسـبه باشـد کاملاً گمراه کننده است.

---

[1]- KeyWord
[2]- Counts

چگالی کلمات کلیدی توسط تساویِ۲-۱۰ تعریف می گردد [FAL95]:

$$\text{چگالی کلمات کلیدی} = KD_i = \frac{tf_i}{L_i} \quad (۲-۱۰)$$

که همانند تساویِ۲-۹ tf_i تعداد دفعاتی که کلمهٔ i اُم در یک سند یافت می شود و L_i تعداد کلمات یک سند می باشد.به عبارتی دیگرچگالی کلمهٔ کلیدی فقط نسبت کلمهٔ کلیدی به کلّ کلمات است.

این نسبت تمرکز کلمات در یک سند را بیام می کند بنابراین چگالی کلمهٔ کلیدی یک سندِ۵۰۰ کلمه ای که در آن کلمهٔ "ایران" ۵ بار تکرار شده است برابر است باKDi=5/500=0.01=1%.

باید توجه داشت که این مقدار در مورد موقعیت نسبی و نیز پراکندگی نسبی کلمه را در سند هیچ توضیحی نمی دهد. این مؤلفه ها ارتباطِ سند و نیز سمانتیکِ موضوعی را تحت تأثیر قرار می دهند.

۴-۱-۳-۲- نارسائیهای چگالی کلمه

تساویِ۲-۱۰ هیچ توضیحی در مورد *وزن معنایی*[1]کلمات در ارتباط با کلمات دیگر در یک سند و یا در مجموعهٔ اسناد نمی دهد.در واقع متخصصان بهینه سازی موتور جستجو[2]و نشان گذاری متور جستجو[3]که وقت خود را بر روی تنظیم چگالی کلمه پس از خرید آخرین آنالیزگرهای چگالی کلمه و فن های وزن کلمات می گذارند، پول و زمان خود را بیهوده تلف می کنند [FAL95].

بر طبق تساویِ۲-۱۰، کلمهٔ k1 که در دو سند متفاوت با اندازهٔ یکسان، بطور مساوی تکرار شده است، صرفنظر از محتوای سند یا طبیعت پایگاه داده، بایستی در دو سند چگالی کلمهٔ یکسان داشته با شد. با این حال اگر فرض شود که چگالی کلمه معیاری برای وزن کلمه قلمداد می شود آنگاه

۱. حجم برشی[4]اطلاعاتی که کلمهٔ پرس و جو شده بازیابی میکند را در نظر نمی گیریم.
۲. بدون ملاحظهٔ مرتبط بودن کلمه به کلمات وزن نمی دهیم.
۳. بدون در نظر گرفتن طبیعت پایگاه دادهٔ جستجو شده، وزنی اختصاص نمی دهیم.

نکات ۱ تا ۳ در بالا مدلِ سالتون را نقض می کند. برطبق تساوی۲-۹ وزن کلمات، نسبت کلمات محلی جدا از پایگاه دادهٔ مورد جستجو نیست و اغلباً یک کلمه مانندK1 که بطور مساوی در دو سندِ متفاوت با اندازهٔ

[1]- Semantic Weight
[2]- Search Engine Optimizer
[3]- Search Engine Marketer
[4]- Shear Volume

یکسان(بدون در نظر گرفتن محتوا) تکرار شده است، بطور متفاوت در یک پایگاه دادهٔ مورد جستجو وزن دهـی می شود [FAL95].

۴-۲- نظریه ی مرکزیت

۱-۴-۲- شهود اصلی تئوری 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 می شود. انسجام و برجستگی در یک سگمنت از گفتار مورد نظر می باشد. همانطور که در ادامه نیز خواهیم دید، تمامی تحقیقات انجام شده بر روی نظریه ی 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 بحث های 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 می دهد و مثلاً در مورد ارتباط معنایی میان جمله ی دوم و بیستم متن ادعایی ندارد.

از آنجایی که بسیاری از تلاش ها بر روی نظریه ی مرکزیت حول مباحث تئوریک آن بوده و کاربردهای تجربی آن کمتر مورد تحقیق قرار گرفته است، می توان ادعا نمود که بسیاری از توانایی ها بالقوه ی این نظریه مورد 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 ساده تر می باشد، ارائه دهد.

در نتیجه یک تئوری بسیار متفاوت با آنچه به طور معمول در زبان شناسی محاسباتی وجود دارد به دست می آید. نکته ی نگران کننده ی خاص در بسیاری از تئوری های زبان شناسی محاسباتی این حقیقت است که، مقالات مرکزیت مانند [GRO95] هیچ الگوریتمی برای محاسبه ی مفاهیمی مانند جمله[۱]و جمله ی قبلی[۲] و رتبه بندی[۳]و شناسایی[۴]که نقش مهمی در این تئوری بازی می کنند را ارائه نکرده اند. محققینی که بر

---

[1] - utterance
[2] - previous utterance
[3] - ranking
[4] - realization

روی مرکزیت کار می کنند، می گویند تا زمانی که این مفاهیم نقش مرکزی در تئوری های انسجام و مرکزیت بازی می کنند، بهتر است که ویژگی های مشخص آن ها برای تحقیقات متوالی دست نخورده باقی بماند. در واقع تعدادی از این مفاهیم مانند رتبه بندی با یک روش متفاوت برای هر زبان تعریف می شوند [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 زبان های مختلف است [KAM98]، [WAL94]، [DIE98]، [STR99] و [TUR98]. نسخه های رقابتی تئوری مرکزیت و ادعاهای مطرح در آن معمولاً به وسیله ی مؤلفین مشابهی ارائه شده است، مثلاً تعاریف مختلف برای CB در [GRO83]، [GRO95] و [GOR93] قابل مشاهده می باشد. در نتیجه محققی که می خواهد، تا یک پیش بینی در مورد مرکزیت را تست کند، یا از مرکزیت برای کاربردهای تمرینی استفاده کند، با تعداد زیادی از نمونه های 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 می شویم، زمانی که تعداد زیادی تعاریف متفاوت برای مفهوم مشابه دستور ارائه می شود - اما باعث می شود که مرکزیت با تئوری هایی که در زبان شناسی محاسباتی با آن روبرو می شویم، نسبتاً متفاوت باشد.

این کمبود جزئیات الگوریتمی نباید منجر به این نتیجه شود که مرکزیت فقط یک تلاش برای تعیین یک زمینه ی تحقیقاتی است، بدون اینکه ادعای خاصی را ارائه کند. در یک تضاد این تئوری ادعاهای بسیار قوی را ایجاد می کند که به زودی در مورد آن بحث می شود. 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 بینی های متفاوتی را ایجاد کنند، درست مثل راه های مختلف برای مشخص کردن دستور که ممکن است نتیجه ی آن، پیش بینی های متفاوت درباره ی نقض محدود سازی یا محدوده ی محتمل متفاوت برای خواندن یک جمله می باشد [REI83]، [SZA97]، [MAY77] و [MAY85]. این آزادی در پر کردن خلأ ها الهام بخش محققانی است که، خود را وقف تهیه ی چنین تعاریفی، برای زبان های خاص یا به طور عمومی کرده اند، آنقدر که فریم ورک های مفهومی که از روی مقالات اخیر بر روی مرکزیت تولید شده اند، پایه ی 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 نمونه های رقابتی، تعدادی از محققین تردید شدیدی در مورد وضعیت تجربی این تئوری دارند: در مورد اینکه کدام ادعاها می توانند توسط ملاک های تجربی پشتیبانی شوند، و اینکه محققین چگونه می توانند پارامترهای ویژه را تحت تأثیر قرار دهند. برای اینکه این مقایسه با معنی باشد باید از مجموعه داده های مشابهی استفاده کرد.

۲-۴-۲- تئوری 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 می گیرد. برای جزئیات بیشتر می توان به [GRO95] و [WAL98] مراجعه کرد.

۳-۴-۲- اصطلاح شناسی و تعاریف

۱-۳-۴-۲- تمرکز محلی، CF و جمله ها

یک فرض بنیادی که تحت لوای مرکزیت وجود دارد، این است که پردازش یک گفتار دربرگیرنده ی بروزرسانی های متوالی وضعیت های محلی است، که به آن تمرکز محلی گویند. تمرکز محلی شامل مجموعه ای به نام CF می باشد که متناظر با کانون پتانسیل گفتار در [SID79] است و می تواند به صورت موجودیت های گفتار دیده شود [KAR76]، [WEB78]، [HEI82] و [KAM93]. تمرکز محلی همچنین شامل اطلاعاتی درباره ی برجستگی های مرتبط یا رتبه بندی موجودیت های CF است. این تمرکز محلی بعد از هر جمله بروز می شود: در این بروزرسانی CF فعلی با CF جدید و تغییرات CF جایگزین می شود.[۱]مجموعه ی CF که در تمرکز محلی با جمله ی U_i در سگمنت گفتار DS تعریف می شود به صورت $CF(U_i, DS)$ دیده می شود، که به طور کلی با مخفف $CF(U_i)$ نمایش داده می شود. ارتباط بین جمله ها و CF[۲]به وسیله ی یکی از شرایط[۳]معروف آنها فرموله می شود[۴][BRE87].

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

۲-۳-۴-۲- رتبه بندی CP و CB

اکنون دو ادعای مهم این تئوری را خاطر نشان می کنیم: یک این که لیست CF رتبه بندی شده است و دوم این که به خاطر این رتبه بندی بعضی از CF ها به برجستگی ویژه ای دست می یابند. تابع رتبه بندی تنها می تواند 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[۵]شناخته می شود. رتبه بندی همچنین در ساخت

---

[۱] - این فرضیه که پردازش گفتمان شامل بروز رسانی های ادامه دار مدل گفتمان است به عنوان قلب تئوری های معروف به گفتمان معنایی پویا نیز قرار دارد [KAR76]، [KAM93] و [HEI82].

[2] - 'Forward-Looking Center'

[3] - 'Constraints'

[۴] - ترتیب مربوط به نمایش قوانین و محدودیت های مربوط به مرکزیت در این گردآوری با نمایش معمول آن ها 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 شود. این سه محدودیت وضعیت یکسانی ندارند: محدودیت دو می تواند به عنوان فیلتر حاکم بر متغیر های مشخص ... دیده شود، محدودیت سه یک تعریف است و محدودیت یک یک ادعای تجربی می باشد.

[5] - 'Predicator-Looking Center'

یکی از CF ها که به CB مشهور است، استفاده می شود. CB در حقیقت نزدیک ترین مفهوم در مرکزیت به اندیشه ی عرفی و کلی موجود در موضوع متن [SGA67]، [CHA76]، [SAN81]، [AND83]، [GIV83]، [REI85]، [VAL90] و [GUN93] می باشد و نقش مرکزی و اصلی در کلیه ی ادعاهای این تئوری در مورد انسجام و برجستگی را بازی می کند. اگرچه در مقاله ی اصلی مرکزیت [GRO83] CB 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): $\mathrm{CB}(U_{\mathrm{i}})$ جمله ی U_{i} است که بالا رتبه ترین عضو $\mathrm{CF}(U_{\mathrm{i}-1})$ است که در U_{i} قابل شناسایی می باشد [BRE87].

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

۴-۴-۲- گذارها

این فرضیه که می گوید گفتمان ها در صورتی که در قالب چنین روشی به صورت عبارت های پی در پی بسته بندی شوند، بسیار منسجم تر از زمانی که، عبارات با موجودیت گفتار منحصر بفرد شروع می شوند، 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[1](بروز سانی) آنها فرموله می شود که از طریق تمرکز محلی استنتاج می شود. تعدادی از این کلاسه بندی های 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 می شود.

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

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

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

[1] - Transition

در ادامه چندین شمای متفاوت از کلاسه بندی ها مورد بررس قرار گرفته است. سپس در مورد اینکه چگونه از این کلاسه بندی ها برای فرموله کردن ادعاهای اصلی تئوری مرکزیت - قانون دو- استفاده می شود، بحث می کنیم.

۵-۴-۲- ادعاهای اصلی

طبق گفته ی [GRO95] بنیادی ترین ادعا در تئوری مرکزیت به این شرح است: " به میزانی که یک گفتار به قیدهای مرکزیت پایبند باشد، انسجام آن افزایش می یابد و بار استنتاجی که به شنونده وارد می شود کاهش می یابد" [GRO95]. آنها ۷ عدد از این قیدها را لیست کرده اند، که ۳ تای آن مستقیماً قابل ارزیابی می باشند. اگرچه در اینجا به دنبال تحلیل تفاوت میان قید و قانون که در [BRE87] آمده است نیستیم، اما می خواهیم از اسامی پیشنهادی [BRE87] برای این ۳ ادعا استفاده کنیم. که در حال حاضر با این اسامی شناخته شده تر هستند:

قید یک (Strong): همه ی 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 ادامه ها بر یک توالی از حفظ ها ارجحیت دارد و به همین ترتیب یک توالی از حفظ ها بر یک توالی از شیفت ها ارجح می باشد.

۶-۴-۲- انتقال های معنایی نظریه مرکزیت

بر اساس پارامترهای تعریف شده در بخش ۲-۳-۴، انتقال های تعریف شده در بخش ۲-۴-۴ و قوانین و قیدهای تعریف شده در ۲-۵-۴، انتقال های معنایی چهارگانه نظریه مرکزیت که بر اساس آنها، جریان انتقال معنا در متن تشخیص داده می شوند، تعریف می شوند. این انتقال ها در شکل ۲-۳ مشاهده می شوند. طبق ادعای [STR99] این چهار انتقال، مورد قبول قریب به اتفاق محققین در این زمینه قرار دارد.

	$CB(U_n) = CB(U_{n-1})$ or $CB(U_{n-1}) = NIL$	$CB(U_n) \neq CB(U_{n-1})$
$CB(U_n) = CP(U_n)$	CONTINUE	SMOOTH-SHIFT
$CB(U_n) \neq CP(U_n)$	RETAIN	ROUGH-SHIFT

شکل ۲-۳- *شمای تصفیه شده از گذارهای مرکزیت*

۵–۲– پیش پردازش های زبانی

برای اجرای روش ها و الگوریتم های مختلف اعم از وزن دهی و نظریه مرکزیت بر روی مجموعه ای از داده های متنی علاوه بر تغییر فرمت داده ها به یکسری متون ساده، احتیاج به تعدادی مراحل پیش پردازشی نیز می باشد. در این بخش، تعدادی از مراحل پیش پردازش زبانی لازم برای آماده سازی این داده ها، مورد بررسی قرار خواهند گرفت. لازم به ذکر است، که این مراحل پیش پردازشی عمومی نیستند و تنها جهت اجرای الگوریتم پیشنهادی این تحقیق، مورد نیاز هستند.

۱–۵–۲– تشخیص زنجیره های مرجعیتی

در زبانشناسی یک هم ارجاعی زمانی رخ می دهد، که چندین عبارت در یک جمله یا متن به عبارت مشابهی اشاره می کنند. در زبانشناسی محاسباتی جایگزینی مرجع از نقش مهمی برخوردار است. به این دلیل که جایگزینی یک ضمیر یا یک عبارت اسمی با مرجع اسمی آن باعث به وجود آمدن درک و تفسیر درست و شفاف تری از متن و از بین رفتن ابهامات می شود. در عین حال با مشخص شدن مرجع بسیاری از فاعل ها، مفعول ها و ... کشف روابط معنایی میان جملات ساده تر خواهد شد.

نکته ی مهم در اینجا بیان تفاوت میان تعریف جایگزینی ارجاعات با جایگزینی ارجاع معنادار[1]می باشد. جایگزینی ارجاع معنادار شامل روالی است که در آن رابطه ی میان عبارت B و عبارت A که در آن عبارت B به صورت معنایی به عبارت A وابسته است، کشف می شود. یک رابطه ی ارجاع معنایی به این معنی است که عبارت A توصیف و تفسیری برای عبارت B محسوب می شود. به عنوان مثال در (1) داریم :

1. The boy entered *the room*. *The door* closed automatically.

در این جمله رابطه ی میان door و room یک رابطه ی ارجاع معنایی می باشد. در حقیقت جایگزینی مرجع زیر شاخه ی جایگزینی ارجاع معنادار می باشد.

تاکنون روش های مختلفی برای تشخیص هم ارجاعی و ارجاع معنادار ارائه شده است که می توان آن ها را به دو دسته ی روش های زبان شناسی و رو ش های یادگیری ماشینی تقسیم کرد.

از مهم ترین روش های زبانشناسی می توان به الگوریتم Hobb’s، نظریه مرکزیت، روش Strube در تشخیص مرجع ضمایر و الگوریتم Bridging Reference اشاره کرد. از شاخص ترین روش های یادگیری ماشینی نیز می توان به Naive Bayes، درخت های تصمیم گیری، فیلدهای شرطی اتفاقی، هم مرجعی از

[1] -Anaphora

طریق خوشه بندی، تشخیص هم مرجعی توسط یادگیری همکارانه، روش های مبتنی بر مجموعه داده ها اشاره کرد [ELA05].

۲-۵-۲- برچسب زنی معنایی نقش کلمات

برچسب زنی معنایی نقش کلمات یکی از مهمترین ابزارهای پیش پردازشی برای اکثر کاربردهای پردازش متن مانند خلاصه سازی، ترجمه ی ماشینی، استخراج مفاهیم، سیستم های مکالمه ی معنایی، سیستم های پاسخگو، متن کاوی و ... می باشد. این ابزار به دنبال تعیین نقش معنایی هر عنصر جمله در ارتباط با فعل آن جمله می باشد. در حین انجام این عملیات نقش گرامری آن عنصر از قبیل فاعل بودن، مفعول بودن چه مستقیم و چه غیر مستقیم، انواع قیود و ... نیز مشخص می گردد. دو روش کلی برای انجام این عملیات وجود دارد که شامل الگوریتم های با ناظر[1]و بدون ناظر[2]می باشد. از روش های با ناظر می توان به [GIL02] اشاره کرد. این روش ها از پیکره های برچسب زده شده ی انسانی به عنوان مجموعه ی یادگیری استفاده می کنند. از جمله ی این پیکره ها FrameNet و PropBank می باشند. از روش های دسته دوم می توان به [SWI04] اشاره کرد.

۳-۵-۲- برچسب زنی نحوی لغات

برچسب گذاری اجزای واژگانی کلام عمل انتساب برچسب های واژگانی به کلمات و نشانه های تشکیل دهنده یک متن است، به صورتی که این برچسب ها نشان دهنده نقش کلمات و نشانه ها در جلمه باشند. درصد بالایی از کلمات از نقطه نظر برچسب واژگانی دارای ابهام هستند، زیرا کلمات در جایگاه های مختلف برچسب های واژگانی متفاوت دارند. بنابراین برچسب گذاری واژگانی عمل ابهام زدایی از برچسب ها با توجه به زمینه [48]مورد نظر است. برچسب گذاری واژگانی عملی اساسی برای بسیاری از حوزه های دیگر پردازش زبان طبیعی از قبیل ترجمه ماشینی، خطایاب و تبدیل متن به گفتار می باشد.

اگرچه تعداد زیادی معتقد نیستند که این کار جزئی از متن کاوی است ؛ ولی برای مثال سیستمی به نام GATE [BON02] در دانشگاه شفیلد، در یک کتابخانه ی دیجیتال به این منظور جایگذاری شده است. GATEشامل ابزاراتی است برای برچسب زدن به کلمات. برای مثال این سیستم می تواند در داخل یک متن، نام موقعیتهای جغرافیایی، نام اشخاص و چیزهایی شبیه این را بیابد. به این خاطر این سیستم بیشتر شامل استخراج اطلاعات است تا استخراج دانش. برچسب زنی نحوی لغات، اغلب نقش بزرگی را در پردازش زبانهای طبیعی بازی می کند. درحقیقت این اولین قدم در پردازش زبان طبیعی است و همانطور که خواهیم دید

---

[1] - Supervised
[2] - Unsupervised

پردازش زبان طبیعی یکی از پایه های متن کاوی است. برچسب زنی نحوی لغات برچسب هایی را به کلمات نظیر اسم، فعل و ... اختصاص می دهد.

تاکنون مدل ها و روش های زیادی برای برچسب گذاری در زبان های مختلف استفاده شده است. این روش ها و مدل ها را می توان به دو دسته کلی تقسیم بندی کرد :دسته اول رهیافت های آماری است که از پیکره های برچسب خورده [49]بهره می جویند و دسته دیگر رهیافت های غیرآماری و مبتنی بر قانون [50]است که بر مبنای یادگیری ماشینی و دانش بشری استوارند. بعضی از روش های گزارش شده را ذکر می کنیم: مدل مخفی مارکوف[51]، سیستم های ماکزیمم آنتروپی[52]، برچسب گذاری مبتنی بر تبدیل [53]و سیستم های مبتنی بر حافظه.[54]

۴-۵-۲- حذف کلمات زائد

کلمات ایست لغاتی هستند که در جملات کاربرد ربطی دارند و زیاد استعمال می شوند مثل " اگر"، " و"، "or"، "the". این کلمات علی رغم اینکه بسیار استفاده می شوند اما از لحاظ معنایی دارای اهمیت کمی بوده و به همین دلیل عموما در فعالیت های مربوط به حوزه پردازش زبان طبیعی در فاز پیش پردازش حذف می شوند. برای حذف این کلمات عموما لیستی از این کلمات از پیش تهیه می شود و سپس در صورت رخداد این کلمات در متن، از سند حذف می شوند. برای زبان انگلیسی چندین لیست از این کلمات منتشر شده است که بطور میانگین شامل ۵۰۰ کلمه می باشند.

۵-۵-۲- ریشه یابی کلمات

ریشه یابی به فرآیند کاهش دادن لغات به ریشه های آنها اطلاق می گردد. بنابراین "computer"و"compute"و'computing" به"compute" که ریشهٔ اصلی است کاهش می یابند. تمامی سیستمای بازیابی اطلاعات نوع یکسانی از "ریشه یاب" را مورد استفاده قرار نمی دهند. در انگلیسی معروف ترین ریشه یاب ، الگوریتم ریشه یاب "مارتین پورتر" است [POR80].

در طول سالیان گذشته ،بسیاری مزایا و معایب استفاده از ریشه یابی را متذکر شده اند. به عنوان مثال ؛ هیچ شکی نیست که رشه یابی کردن تضمین می کند که سند هایی که همگی شامل اشتقاقهای متفاوتی از کلمهٔ موجود در پرس و جو هستند، در مجموعه جواب نهایی هستند. ریشه یابی همچنین سایز فایل وارونه را کاهش می دهد. از طرفی دیگر ریشه یابی در حدّ بالا عملی نیست و باعث آزار کاربران می گردد. همانطور کهAvi Rappoport متخصص در امورِ بازیابی اطلاعات می گوید:"در بسیاری از موارد، موتور جستجو در

[1]- Martin Porter's Stemming Algorithm

وهلهٔ اول نمی تواند مطابقتهای دقیق را نشان دهد؛ چون متن ریشه یابی شده در ایندکس ذخیره شـده اسـت، این امرممکن است که باعث ذخیره فضای دیسک شود اما باعث آزار کاربران می گـردد.این مشـکل بیشـتر در مورد افعال رخ می دهد.بهتر و آسانتر آن است که ریشه یابی در کوئری انجام شود وبرای نگارش های مختلـف کلمه جستجو شود تا آنکه تنها نگارشِ ریشه یابی شده ذخیره گردد" [POR80].

ریشه یابی کلمات با توجه به رشد پردازش زبان طبیعی رشد کاربردهای فراوانی پیدا کرده است. به طور کلی دو کاربرد عمده برای ریشه یابی کلمات مرسوم است:

۱. ریشه یابی کلمات در ماشین های مترجم.
۲. ریشه یابی کلمات در سیستم های بازیابی اطلاعات

در سیستمهای بازیابی اطلاعات معمولاً یک پایگاه دادهٔ بسیار بزرگ وجود دارد که بایستی پردازش و بازیابی اطلاعات بر روی آنها صورت گیرد. هرچه شبکه های معنایی استخراج شده از این اطلاعات دقیق تر و گسترده تر باشد، امکان فراهم شدن اطلاعات استخراج شدهٔ بیشتر ، راحتتر است. یکی از کاربردهای ریشه یابی در امکان فراهم کردن شبکه های معنایی گسترده تر در سیستم های پردازش متن و بازیابی اطلاعات است.

کلمات در هر زبان به دو دستهٔ جامد و مرکب تقسیم می شوند .به کلماتی که از دیگر کلمات مشتق شده اند، کلمات مرکب گفته می شود. کلمات جامد کلماتی هستند که در زبان از هیچ کلمه ای مشتق نشده اند. یافتن ریشهٔ کلمات را اصطلاحاً *ریشه یابی کلمات* میگوئیم. به عنوان مثال در زبان فارسی ، "آموزگار "کلمهٔ مرکبی است که از ترکیب "آموز" و "گار" تشکیل شده است یا در زبان انگلیسی ، "teacher" از کلمات "teach" و "er" تشکیل شده است.

۱-۵-۵-۲- الگوریتم های ریشه یابی

تاکنون الگوریتم های زیادی برای ریشه یابی کلمات معرفی شده است. با این حال در یک دسته بندی کلی می توان دو ایدهٔ اصلی در تمامی آنها جستجو کرد:

- **شبکه های کلمات :**در این روشها از ارتباط کلمات با هم در یک شبکهٔ معنایی استفاده می شود. کلماتی که دارای یک ریشه هستند مشخص شده و در یک شبکهٔ گراف مانند به طوریکه اعضای یک دسته تشکیل یک خوشه را داده اند، نگهداری می شود. این روشها نیاز به نظارت بسیار زیاد عامل انسانی دارند ، زیرا که هنوز الگوریتمهایی که با آنها بتوان شبکه های قابل اطمینان یافت

[1]- Keyword Stemming & Word Forums; Search Engine Watch Forums

وجود ندارد. اما در مقابل دامنهٔ کارائی این الگوریتمها بسیار گسترده است و در سیستمهای داده کاوی و متن کاوی کارائی مناسبی دارند .نمونه ای از سیستمهای موفق ریشه یابی با این رویکرد در "*استفادهٔ شبکهٔ کلمه برای بازیابی اطلاعات*"[55]وجود دارد [KRO93].

ریخت شناسی :در رویکرد دومی که به مسألهٔ ریشه یابی وجود دارد، از قوانین ساخت کلمات استفاده می شود. این الگوریتم ها با بررسی روشهای ساخت کلمات، ریشهٔ کلمات را پیدا می کنند. این روشها به علت مشخص بودن روشهای ساخت کلمات امکان مکانیزه شدن خوبی را فراهم کرده اند، اما مشکل اصلی در آنها هنگامی رخ می دهد که دو کلمهٔ جامد با دو تغییر مختلف یک کلمهٔ مرکب یکسان می سازند. به عنوان مثال کلمهٔ "goes"می تواند به معنای goe + s به معنای زنبورهای عسل باشد و یا go + es به معنی صورتی از فعل رفتن .نمونه هایی از الگوریتم های ریخت شناسی الگوریتمهای معروف *پورتر* [56]و الگوریتم *کراوتز* [57] هستند که در ادامه به بررسی آنها می پردازیم [KRO93].

۶-۲- کارهای مرتبط

به طور کلی می توان گفت که هدف از وزن دهی کلمات تبدیل داده های موجود در حوزه ی متن به داده هایی در محیط دیگری است، تا از این طریق بتوان به صورت ساده تری عملیات لازم بر روی متون را انجام داد. تعداد زیادی از روش و توابع وزن دهی به کلمات تاکنون ارائه و تست شده اند [LEE95]، [SAL75]، [SAL83]، [SAL88]، [FUH91]، [LUH57]، [SPA72] و [SPA73]. دیدگاه کلی نسبت به وزن تعیین میزان اهمیت یک کلمه در یک متن می باشد. عمده روشهایی که تاکنون ابداع گردیده اند مبتنی بر معیارهای آماری می باشند و اصلی ترین فاکتورهای دخیل در این معیارها نیز فاکتورهای TF (فرکانس کلمه)، IDF (معکوس فرکانس سند) و LN (نرمال سازی طول) می باشند.

تمام این روش ها با تغییر در خواص آماری یک سند، رفتارهای متفاوتی را از خود بروز می دهند. به عنوان مثال فاکتور TF در اسناد با طول زیاد به خوبی عمل می کند، اما با کوتاه شدن طول سند نتایج بسیار بدی را از خود نمایش می دهد. دیگر فاکتورها نیز به همین نحو نسبت به تغییرات در ویژگی های آماری اسناد، بسیار حساس هستند[MOE00].

از سوی دیگر تمامی این نوع از فاکتورها و به خصوص TF در شناسایی دقیق اهمیت کلمات دچار اشکالات زیادی هستند. فرض کنید دو کلمه ی نسبتاً کم کاربرد اما تخصصی در یک متن استفاده شده اند. یکی از کلمات به لحاظ معنایی با مفهوم کلی متن در ارتباط است، در حالی که کلمه دیگر فقط به لحاظ بیان مطلب مورد استفاده قرار گرفته است. در این حالت تمام روش های آماری این دو کلمه به یک میزان وزن دهی می نمایند [MOE00].

در این بین تلاشهایی نیز در جهت دخالت دادن عوامل معنایی در وزن دهی به کلمات صورت گرفته انـد. ایـن تلاش ها، در یک دسته بندی از نظر احتیاج به دانش اولیه، به دو دسته کلی تقسیم می شوند.

۱. روش های نیازمند به دانش اولیه:

در این دسته، روش وزن دهی ممکن است به دو دلیل احتیاج به دانش اولیـه داشـته باشـد. روش های وزن دهی در این دسته، یا از طریق دانش اولیـه بـه غنـی سـازی مجموعـه داده اولیه و بسط مجموعه کلمات به مجموعه بزرگتری از کلمات می پردازنـد و یـا روابـط میـان کلمات را از طریق جستجوی این کلمات در مجموعه دانش اولیه استخراج می کنند. یکی از اولین نمونه های این دسته کـاری اسـت کـه در [MOR88] و [MOR91] انجـام شـده است. در این پژوهش از زنجیره های لغوی به عنوان محور اصلی شناسایی میزان ارتباط یک کلمه با محتوای معنایی متن استفاده نموده انـد. یـک زنجیـره ی لغـوی در اصـل بـه بیـان ارتباطات معنایی لغات در چهار دسته یکسان، هم معنی، ابر معنی و زیر معنی مـی پـردازد. روش تعیین این ارتباطات استفاده از یک فرهنـگ لغـات دارای لینـک بـین کلمـات[1]اسـت. [MOR91] از Roget به عنوان چنین فرهنگ لغتی در کار خود استفاده کرده اند. البتـه فرهنگ لغتـی ماننـد Roget امکـان خودکارسـازی انجـام فرآینـدها را نـدارد (بـر خـلاف WordNet). از سوی دیگر [HIR98] برای خودکار سازی این فرآینـد از wordnet بـه عنوان فرهنگ لغت برخط استفاده کردند. با این تفاوت کـه بـه علـت تمـایز در نـوع روابـط تعریف شده در آن با قوانین تشکیل زنجیره های لغوی، این قوانین را تغییر داده و فقـط بـر 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 همـین روش با wordnet، یک الگوریتم برچسب زنی نحوی، پارسر کم عمـق[2]و یـک الگـوریتم قسـمت بندی به نتایج خوبی دست یافتند.

در [KAN05] نیز از زنجیره های لغوی برای خوشه بنـدی مـتن و سـپس وزن دهـی بـه لغات استفاده شده است. روش کار به این صورت است کـه ابتـدا بـا تشـکیل زنجیـره هـای 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 روابـط بـین لغـات، هـر خوشـه بـا

[1] - Thesaurus
[2] - Shallow parser

روشی دقیقاً مشابه با رتبه بندی صفحات[1] 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 80% کاهش حجم در نگهداری اندیس صفحات دارد.

همچنین در [BAR09] 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[2] 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 [BAR09] 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.

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.

---

[1] - Page rank
[2] - Semantic kernel

۲. روش های بدون نیاز به دانش اولیه:

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

از این دست روش ها می توان به کاری که در [ERN07] ارائه شده است، اشـاره کـرد. در این تحقیق به کلمات به عنوان افرادی از یـک شـبکه ی اجتمـاعی نگریسـته شـده اسـت 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.

$$S(w) = (1-\lambda)d_w + \lambda \sum_{u \in r(w)} S(u)C_{w,u} \qquad (۲-۱۱)$$

در رابطه ی بالا d_w وزن اولیه هر کلمه اسـت کـه بـا اسـتفاده از معیـاری ماننـد TF-IDF مصاحبه می شود و در یک مدل یادگیری سعی می شود تا $C_{w,u}$، یعنـی وزن ارتبـاط بـین 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 زمینـه[1]مشـترکی قـرار گرفتـه باشند. این زمینه ی مشترک عبارت است از یک قطعه متن با اندازه ی محدود کـه پیوسـته باشد. به عنوان مثال یک لینک و یا یک پاراگراف و یا کلماتی که با فاصله مشخصـی از یـک کلمه قرار دارند تشکیل یک زمینه را می دهند. در نهایت محاسـبه C(w,u) بـا اسـتفاده از رابطه ۲-۱۲ و ۲-۱۳ انجام می شود.

$$C_{w,u} = \sum_{\hat{w} \in occ(w)} \sum_{\hat{u} \in ict(\hat{w},u)} C_p(\hat{w}, \hat{u}) \qquad (۲-۱۲)$$

به طوریکه

$$= \prod_{i=0}^{i=n} \sigma(\alpha_i x_i + \beta_i) C_p(\hat{w}, \hat{u}) \qquad (۲-۱۳)$$

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[2]تحت رابطه ۲-۱۴ می باشد.

$$\sigma(x) = {1}/{(1 + e^{-x})} \qquad (۲-۱۴)$$

---

[1] - Context

[2] - Sigmoid

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[1] 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 مقادی CB جمله مشخص شده است. سپس با استفاده از یک مدل LDA[2] مبتنی بر ویژگی های بدست آمده یک کاربرد خاص، که خلاصه سازی متون می باشد مورد پیگیری قرار گرفته است.

عمده ی روش هایی که به داخل کردن ویژگی های معنایی و زبانی در وزن دهی به کلمات پرداخته اند، در دسته اول قرار می گیرند. از دسته دوم روش ها وزن دهی معنایی، نمونه های کمی موجود می باشد، که از میان آنها، به دو روش مؤفق تر اشاره شد. روش های دسته اول، علیرغم بهبود دقت کاربردهایی که از وزن دهی کلمات استفاده می کنند، چالش های جدیدی را ایجاد کرده اند. عمده ی این چالش های جدید عبارتند از:

۱. استفاده از فرهنگ های لغت و یا مجموعه دانش اولیه مانند آنتولوژی های واژگان. این وابستگی خود می تواند منجر به انتشار خطا در محاسبه وزن شود.
۲. بالا بردن بی رویه مرتبه زمانی محاسبه وزن به دلیل احتیاج به پردازش های فراوان بر روی مجموعه دانش های اولیه و استفاده از انواع روش های یادگیری
۳. عدم وفاداری به مجموعه داده ی اولیه و تغییر و بسط و گسترش مجموعه کلمات به انحاء مختلف.
۴. وابسته بودن دقت به حجم دانش اولیه

در عین حال این روش ها مشکلات عمده موجود در روش های آماری را تا حد زیادی حل کرده اند. این مشکلات عبارتند از:

۱. عدم توجه به معنا
۲. عدم توجه به ویژگی های زبانی

[1] - Nueral Network
[2] - Latent Dirichlet Allocation

۳. عدم توجه به ویژگی های فرازبانی
۴. پایین بودن دقت
۵. عدم توانایی در خوشه بندی اولیه کلمات

البته تقریباً تمامی روش های وزن دهی کنونی، حداکثر مشکلات ۱، ۲، ۴ و ۵ از مشکلات پنج گانه روش های بالا را حل کرده اند. مشکل شماره ۳ به دلیل ناکافی بودن سطح دانش فعلی، از تلاش های زیادی بـر روی آن برخوردار نبوده است.

در جدول ۵-۲ تعدادی از روش های بالا که بیشترین تعداد ارجاع در مقالات دیگر به آنها وجود داشـته اسـت، بر اساس چهار ویژگی رویکرد وزن دهی، دانش اولیه مورد استفاده، افزایش دقت، روش وزن دهی آماری مبنـا، حوزه کاربرد، مشکلات حل شده و مشکلات پیش رو آنها مورد مقایسه قرار گرفته اند. از پیاده سازی و مقایسـه این روش ها در این پایان نامه، به دلیل استفاده آنها از ابزارهای خاص و شخصی، معذور هستیم.

جدول ۲-۵- مقایسه روش های معنایی وزن دهی به کلمات بر اساس ۷ ویژگی اصلی

روش وزن دهی	رویکرد وزن دهی	دانش اولیه	افزایش دقت	روش وزن دهی آماری مبنا	حوزه کاربرد	چالش های پیش رو	مشکلات حل شده
[HIR98]	زبانی، معنایی	wordnet	--	--	تشخیص دسته های کلمات بر روی داده های برگرفته شده از خبرهای Wallstreet journal	چالش ۱، ۲ و ۴	مشکل ۱، ۲، ۴ و ۵
[CAR02]	زبانی، معنایی	wordnet	میانگین ۱۰٪ نسبت به TF-IDF	--	بازیابی اطلاعات بر روی داده های MED	چالش ۱، ۲ و ۴	مشکل ۱، ۲، ۴ و ۵
[KAN05]	زبانی، معنایی	wordnet	میانگین ۱۵٪ نسبت به TF-IDF	TF	بازیابی اطلاعات بر روی داده های TREC2	چالش های چهارگانه	مشکل ۱، ۲، ۴ و ۵

[BAR09]	معنایی	wordnet	--	--	TF-IDF	کلاس بندی بر روی داده های Reuters-10	چالش های چهارگانه	مشکل ۱، ۴ و ۵
[CHA08]	معنایی	--	--	--	خوشه بندی اسناد بر روی داده های TREC2008	چالش های چهارگانه	مشکل ۱، ۴ و ۵
[WAN08]	معنایی	wikipedia	میانگین ۳٪ نسبت به TF-IDF	TF-IDF	کلاس بندی بر روی داده های Reuters-21578	چالش های چهارگانه	مشکل ۱، ۴ و ۵
[BLO06]	معنایی	wordnet	میانگین ۱۰٪ نسبت به TF-IDF	--	بازیابی اطلاعات بر روی داده های MED	چالش های چهارگانه	مشکل ۱، ۴ و ۵
[LUO11]	زبانی، معنایی	wordnet	میانگین ۴٪ نسبت به TF-IDF	TF	کلاس بندی داده های WebKB و Rueters-21578	چالش های چهارگانه	مشکل ۱، ۲، ۴ و ۵
[WIT09]	معنایی	wordnet	--	--	کلاس بندی داده های Reuters-21578	چالش های چهارگانه	مشکل ۱، ۴ و ۵
[ERN07]	معنایی	--	میانگین ۱۷٪ نسبت به TF-IDF	--	Question Answering بر روی داده های WebCrow	چالش ۲	مشکل ۱، ۴ و ۵
[DAS09]	زبانی	--	--	--	خلاصه سازی بر روی داده های DUC2005، TAC2008 و	ندارد	مشکل ۱، ۲، ۴ و ۵

		TAC2009					

۷-۲- خلاصه فصل

در این فصل مروری کلی بر مفاهیم دخیل در وزن دهی به کلمات و به خصوص روش پیشنهادی داشتیم. ابتدا به بررسی و بازگویی مفهوم بازیابی اطلاعات و مفاهیم مرتبط با آن پرداختیم. گام های ۵ گانه موجود در یک سیستم بازیابی اطلاعات را برشمردیم و سپس جایگاه وزن دهی به کلمات به عنوان یکی از مهمترین گام های موجود در این سیستم ها را نشان دادیم. سپس به بررسی و مطالعه ی اجمالی مفهوم وزن دهی به کلمات پرداختیم. مفاهیم پایه ای مانند وزن محلی و وزن سراسری، نمایش چندین معیار مشهور از هر کدام، معیارهای نرمال سازی وزن و معیارهای ارزیابی سیستم های بازیابی اطلاعات و روش های ون دهی به کلمات مورد بررسی قرار گرفت.

در ادامه و به عنوان پایه ی رویکرد معناگرا در روش پیشنهادی وزن دهی به کلمات، به معرفی نظریه مرکزیت پرداختیم. ادعای جدی این نظریه در تشخیص انسجام و برجستگی محلی بر اساس تعریف چندین پارامتر بر روی جملات یک متن و کلمات آن جملات با توجه به نقش گرامری آنها می باشد. بر اساس این تعاریف ۴ سری از انتقال های معنایی تعریف می شوند و بین هر دو جمله یک انتقال معنایی محاسبه می شود.

در خاتمه به بررسی تعدادی از روش های موجود وزن دهی به کلمات که دیدی معناگرا دارند، پرداختیم. همانطور که مشاهده شد، این روش ها چالش بزرگ روش های وزن دهی آماری که عدم توجه به معنا و ویژگی های زبانی است، و باعث به وجود آمدن دقت پایین در کاربردهای زبانی می شود را تا حد مناسبی حل کرده اند. اما خود باعث به وجود آمدن مشکلات سه گانه ای شده اند. این مشکلات استفاده از فرهنگ های لغت و یا مجموعه دانش اولیه مانند آنتولوژی های واژگان، بالا بردن بی رویه مرتبه زمانی محاسبه وزن به دلیل احتیاج به پردازش های فراوان بر روی مجموعه دانش های اولیه و استفاده از انواع روش های یادگیری و عدم وفاداری به مجموعه داده ی اولیه و تغییر و بسط و گسترش مجموعه کلمات به انحاء مختلف می باشند.

فصل سوم :

روش پیشنهادی

۳- روش پیشنهادی

۳-۱- مقدمه

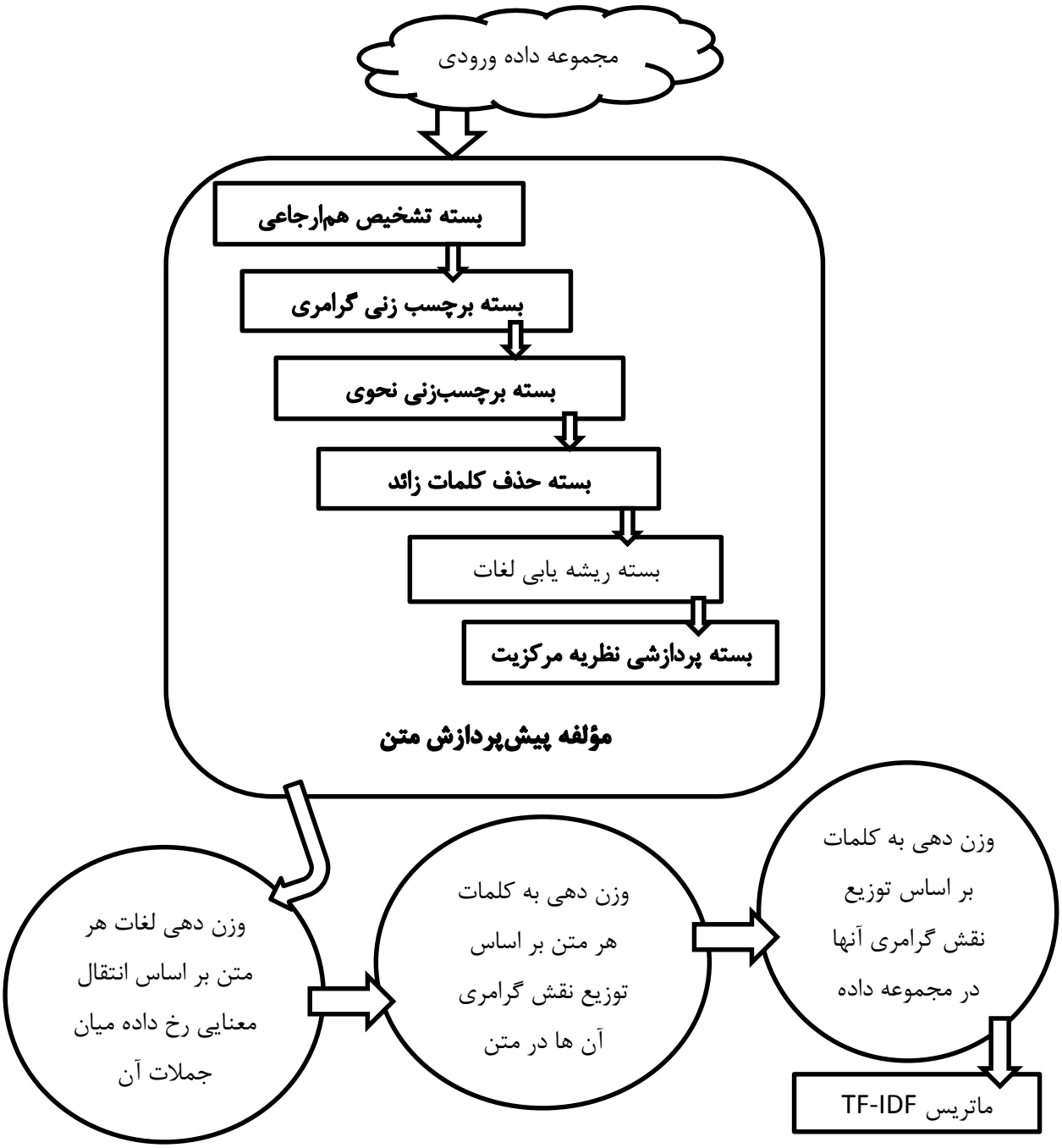
وزن دهی به کلمات به عنوان ویژگی های یک بردار سند از اهمیت بالایی در بازیابی اطلاعات از داخل متن برخوردار است. علت این امر آن است که با یک معیار وزن دهی مناسب می توان تا حد زیادی مراحل درک معنا و استخراج مفهوم یک سند را تسهیل بخشید. همانطور که دیدیم، در حال حاضر اکثر معیارهای مطرح بر پایه ی اطلاعات آماری سند محاسبه می شوند. خلأ استفاده از ویژگی های زبانی سند در محاسبه ی وزن کلمات، از بروزترین و مورد توجه ترین چالش های امروزی محققان در زمینه ی بازیابی اطلاعات متن می باشد. علت این توجه در این است که مفاهیم موجود در یک سند توسط کلمات موجود در آن سند حمل شده و استباط می گردد. ویژگی های زبانی کلمات، به خصوص نقش گرامری آنها از جمله پارامترهایی هستند، که قابلیت تعریف نقش معنایی کلمه در یک سند را دارند. از این رو در روش پیشنهادی در این تحقیق، توجه شایانی به پایه گذاری معیار وزن دهی بر مبنای نقش های گرامری کلمات در متن و سایر ویژگی های زبانی آنها شده است.

۳-۲- ساختار سیستم پیشنهادی

در این فصل روش پیشنهادی برای وزن دهی به کلمات در ماتریس کلمه-سند ارائه خواهد شد. این روش برای وزن دهی به بردار نمایش سند استفاده می شود. بنابراین در یک کاربرد بازیابی اطلاعات پیش از مقایسه ی پرس و جوی هدف با بردارهای اسناد، ابتدا باید وزن هر کلمه در هر بردار سند توسط روش پیشنهادی محاسبه گردد. شکل ۳-۱ نحوه ی گردش کار سیستم پیشنهادی را نشان می دهد.

برای نمایش یک پرس و جو، حالت ساده و ثابتی در تمامی کار مورد نظر است، که شامل یک رشته از کلمات می باشد. برای وزن دهی به اسناد ابتدا مراحل پردازشی بر روی هر سند صورت می گیرد. پس از این مرحله ی پردازشی ماتریس کلمه-سند تشکیل شده و در بانک اسناد ذخیره می شود. این مراحل پردازشی به تفکیک و تفصیل در بخش ۳-۳ مورد بحث و بررسی قرار خواهد گرفت. سپس الگوریتم وزن دهی به کلمات بر روی هر بردار سند از ماتریس کلمه-سند اعمال می گردد. این الگوریتم ترکیبی از حدس های ذهنی و پارامترهای الگوریتمی می باشد که به تفکیک و تفصیل در بخش ۳-۴ ارائه خواهد شد. پارامترهای زبانی شامل شرایط و محدودیت های مندرج در نظریه ی مرکزیت است. نظریه ی مرکزیت، یک تئوری زبانی در مورد نحوه ی انتقال معنا در یک توالی از جملات می باشد. از این نظریه برای انتخاب و برجسته نمودن تعدادی از جملات یک متن و اضافه کردن یک ضریب در محاسبه ی وزن کلمات این جملات استفاده می کنیم. یکی از چالش های پیشرو در استفاده از این نظریه، توجه معیار مطرح در آن به انسجام و برجستگی به

صورت محلی است. در حالیکه برای تشخیص و استخراج مفاهیم از میان اسناد احتیاج به یک معیار وزن دهی به کلمات است، که توجه به اهمیت سراسری کلمات نیز داشته باشد. در نتیجه در بخش ۳-۴ نشان خواهیم داد که چگونه معیارهای زبانی با تغییر یک معیار وزن دهی خوب به نام TF-IDF به بهبود نتایج کمک خواهند کرد. حدس های ذهنی نیز شامل توجه به مواردی مانند نقش کلمات، تکرار آن ها در نقش های یکسان در سطح یک یا چندین سند می باشد.



شکل ۳-۱- ساختار کلی سیستم وزن دهی پیشنهادی

یک تفاوت عمده میان روش پیشنهادی و روش های ابداعی تاکنون، توجه ویژه به نقش های گرامری اسامی در اسناد می باشد که در حقیقت حمل کننده ی بار معنایی آنها می باشند. این توجه هم به صورت محلی و توسط معیارهای الگوریتمی نظریه ی مرکزیت و هم به صورت سراسری و با توجه به نوع نقش گرامری یک کلمه در متون مختلف انجام پذیرفته است. در بازیابی یک پرس و جو در میان بانک اسناد، تنها ماتریس کلمه-سند که حاوی بردارهای وزن دهی شده ی اسناد می باشد، مورد ارجاع قرار می گیرد.

۳-۳-پیش پردازش های زبانی

در پیاده سازی الگوریتم پیشنهادی برای وزن دهی به کلمات در بردار اسناد احتیاج به دو دسته از پردازش های اولیه زبانی بر روی مجموعه ی اسناد می باشد. یک دسته از این پردازش ها در تمامی کاربردهای بازیابی اطلاعات متنی مشترک است و دسته ی دیگر در برخی از کاربردها از جمله الگوریتم پیشنهادی در این تحقیق مورد نیاز می باشد.

برای پیاده سازی الگوریتم وزن دهی پیشنهادی نیز، احتیاج به تعدادی از پیش پردازش های زبانی می باشد. این پردازش ها که اکثراً برای پیاده سازی و اعمال نظریه ی مرکزیت بر روی مجموعه ی اسناد مورد نیاز هستند، شامل تشخیص زنجیره های مرجعیتی[۱] برچسب زنی معنایی نقش کلمات[۲] برچسب زنی نحوی کلمات[۳] و تعیین انتقال معنایی صورت گرفته میان جملات یک متن[۴] می باشند.

در ادامه برای آماده سازی یک ماتریس از بردارهای اسناد در فضای کلمات موسوم به ماتریس کلمه-سند، احتیاج به پیش پردازش های معمول، معلوم و مشترکی بر روی اسناد به لحاظ زبانی می باشد. این پردازش ها شامل شکست سند به جملات تشکیل دهنده ی آن[۵] حذف کلمات زائد[۶] و ریشه یابی کلمات[۷] می باشد.

این دو مجموعه از پردازش ها به صورت متوالی و بر روی یک مجموعه از داده ها اعمال نمی شوند. به عنوان مثال برای پیش پردازش های گروه اول، تغییر نیافتن شکل کلی اسناد مانند کم شدن کلمات و شکسته شدن متن به جملات تشکیل دهنده ی آن مهم است. چرا که نظریه ی مرکزیت برای تشخیص انسجام و برجستگی محلی در سطح یک سند، احتیاج به یک توالی از جملات دارد. در حالیکه پردازش های دسته ی دوم به صورت جدا از هم و بر روی مجموعه ی اسناد حتی به صورت تغییر یافته نیز قابل اعمال هستند.

---

[1] - Coreference Resolution
[2] - Semantic Role Labeling
[3] - Part of Speech Tagging
[4] - Centering Theory Transition Recognizing
[5] - sentence splitting
[6] - stop words
[7] - Stemming

در الگوریتم پیشنهادی برای کاهش میزان زمان مصرفی مورد نیاز در پیش پردازش های زبانی تمامی این پردازش ها به غیر از برچسب زنی معنایی نقش کلمات و تعیین انتقال معنایی صورت گرفته میان جملات یک متن به صورت موازی انجام شده است.

هر چند که ماهیت مجموعه ی داده ی مورد پردازش در هر دسته به دلیل ساختار آنها متفاوت است، اما می توان با کمی دستکاری، این پردازش ها را به صورت متوالی انجام داده و ماتریس نهایی یکدستی را ایجاد کرد. در ادامه پیش پردازش های مربوط به هر دسته به صورت مشروح مورد بررسی قرار می گیرند. در این بین برای درک درست از پردازش مربوط به نظریه ی مرکزیت، به شرح کامل این نظریه نیز خواهیم پرداخت.

۱-۳-۳- پیش پردازش های مرتبط با الگوریتم پیشنهادی

این دسته از پیش پردازش ها بر روی مجموعه ی اسناد صرفاً به منظور آماده سازی این مجموعه جهت اجرای الگوریتم پیشنهادی بر روی آنها انجام می شوند. جهت اعمال شرایط و محدودیت های نظریه ی مرکزیت برای تشخیص برجستگی و اهمیت محلی کلمات و نیز اندازه گیری تکرار نقش گرامری و معنایی کلمات در کل مجموعه، جهت تشخیص اهمیت سراسری کلمات، به این پردازش ها نیازمند هستیم. مجموعه ی اسناد در کاربردهای مرتبط با زبان معمولاً شامل تعدادی فایل های متنی به فرمت XML و یا در قالب استاندارد TREC[1] می باشند. در ادامه به تشریح هر یک از این پیش پردازش ها پرداخته و فواید و چالش های هر کدام را خاطرنشان خواهیم کرد.

۱-۱-۳-۳- تشخیص زنجیره های مرجعیتی

تا این مرحله، تنها تغییر در مجموعه ی اسناد، حذف و جایگزینی تعدادی از ضمایر و گروه های اسمی با یک گروه اسمی دیگر می باشد. ساختار کلی اسناد که معمولاً به صورت فایل های متنی XML و یا در قالب استاندارد TREC می باشد، تا این مرحله، تغییر نمی یابد.

۲-۱-۳-۳- برچسب زنی معنایی نقش کلمات

خروجی مرحله ی قبل پس از شکستن هر متن به جمله های تشکیل دهنده اش، بر اساس محل قرار گرفتن (.) به عنوان ورودی این مرحله محسوب می شود.

[1] - Text Retrieval Conference: http://trec.nist.gov/

در این مرحله یک ماتریس مبتنی بر نقش های کلمات تشکیل می دهیم. از این پس این ماتریس را جمله-لغت می نامیم. مطابق رابطه ی ۳-2 سطرهای این ماتریس شامل جملات موجود در اسناد و ستون های ماتریس شامل کلمات موجود در متون می باشند. برای هر کلمه نقش گرامری آن نیز نگهداری می شود. نقش های موجود برای یک کلمه که مورد استفاده در این تحقیق می باشند، در رابطه ۳-۱ بر اساس میزان اولویت آنها دیده می شوند. همینطور در ستون آخر ماتریس شماره ی سند هر جمله نیز نگهداری می شود. همچنین برای سادگی کار در مراحل بعدی، ستون های ماتریس، در همین جا بر اساس اهمیت نقش گرامری آنها یعنی فاعل، صفت های مرتبط با فاعل، مفعول غیر مستقیم، صفت های مرتبط با مفعول غیر مستقیم، مفعول مستقیم، صفت های مرتبط با مفعول مستقیم، قید و صفت های مرتبط با قیدها مرتب خواهد شد. همانطور که می بینیم، از این پس ساختار جدید مجموعه ی اسناد، که به شکل ماتریس است را مورد پردازش قرار خواهیم داد.

(۳-۱) فاعل> صفت فاعل> مفعول غیر مستقیم > صفت مفعول غیر مستقیم > مفعول مستقیم > صفت مفعول مستقیم > قید مکان > صفت قید مکان > دیگر

ابعاد این ماتریس m×(n+1) می باشد که n حداکثر تعداد کلمات یک سند در مجموعه ی اسناد و m تعداد کل جملات مجموعه ی اسناد می باشد.

Sentence-Term Matrix:

Term & It's Grammar Role

Order based on priority

Doc NO

$$\text{Sentence}\begin{bmatrix} T_{1,1}, Role & \cdots & \cdots & \cdots & T_{1,n}, Role & D_1 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & T_{i,j}, Role & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ T_{m,1}, Role & \cdots & \cdots & \cdots & T_{m,n}, Role & D_k \end{bmatrix}_{m\times(n+1)} \quad (۳-۲)$$

۳-۱-۳-۳- برچسب زنی نحوی لغات

ورودی ابزار POS Tagger یک متن می باشد. سپس بر اساس خروجی تولید شده، به تصحیح ماتریس جمله-لغت می پردازیم. در این تصحیح تنها ترتیب کلماتی که دارای یک نقش گرامری مانند فاعل هستند، تغییر خواهد کرد. به این ترتیب که از گروه اسمی که شامل یک موصوف و چندین صفت است، موصوف به عنوان کلمه ی با اهمیت تر شناخته خواهد شد. به عنوان مثال در این تصحیح، اگر یک گروه اسمی که فاعل است، شامل یک اسم و تعدادی صفت باشد، پس از اعمال POS Tagging اسم این گروه اسمی به عنوان

فاعل اصلی شناخته می شود. خروجی این مرحله، ماتریس تصحیح شده ی جمله-لغت می باشد که در آن ترتیب کلمات دقیقاً مشخص شده است.

۴-۱-۳-۳- ریشه یابی

در این مرحله ورودی ماتریس جمله-لغت می باشد که یکبار عمل ریشه یابی با استفاده از الگوریتم پورتر بر روی آن انجام شده و تمامی کلمات با ریشه هایشان جایگزین خواهند شد. همینطور کلیه کلمات زائد نیز حذف می گردند.

۵-۱-۳-۳- تشخیص انتقال های معنایی میان جملات

در این بخش پردازش های مرتبط با تشخیص انتقال معنایی صورت گرفته بین هر دو جمله ی متوالی در هر متن را تشریح خواهیم کرد. در حال حاضر همه ی اسناد موجود در مجموعه ی اسناد، به لحاظ زنجیره های ارجاعی شفاف اند. نقش گرامری تمامی اسامی موجود در هر جمله نیز در ارتباط با فعل آن جمله مشخص شده است. همینطور، پس از مشخص شدن نقش گرامری که بعضاً به یک گروه اسمی مرتبط می شود، توسط برچسب زن POS تنها یک اسم به عنوان نماینده و مهم ترین کلمه ی آن گروه اسمی تعیین گردید.

در این قسمت به تشریح نظریه ی مرکزیت و پارامترهای آن می پردازیم و نحوه ی اعمال آن بر روی مجموعه ی اسناد را تشریح خواهیم کرد.

۱-۵-۱-۳-۳- اعمال نظریه مرکزیت و پارامترهای آن بر روی مجموعه ی اسناد

همانطور که مشاهده شد برای اعمال نظریه ی مرکزیت بر روی یک متن، باید پارامترهای اصلی این نظریه که شامل CF، CP و CB می باشند برای هر جمله از متن مورد نظر مشخص شود. در این مرحله ماتریس جمله-لغت رابطه ی ۳-۲ مورد پردازش نظریه ی مرکزیت قرار می گیرد. همانطور که دیدیم این ماتریس، یک مجموعه از داده ها می باشد، که به لحاظ ترتیبی که برای پارامتر CF نظریه ی مرکزیت مورد نیاز است، مرتب می باشد.

حال باید قوانین مربوط به انتقال های معنایی نظریه ی مرکزیت برای هر دو جمله ی متوالی بر روی ماتریس جمله-لغت اعمال گردد. نتیجه ی حاصل از این مرحله ماتریس قبلی می باشد که دو ستون جدید به آن اضافه شده است. ستون اول پارامتر CB برای هر جمله می باشد. و ستون دوم نشان دهنده ی انتقال معنایی صورت گرفته میان جمله ی i و i+1 می باشد. ماتریس جمله-لغت جدید در رابطه ۳-۳ دیده می شود.

Sentence-Term Matrix:

$$\text{Sentence}\begin{bmatrix} T_{1,1}, Role & \cdots & T_{1,j}, Role & \cdots & T_{1,n}, Role & D_1 & CB_1 & Tr_1 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & T_{i,j}, Role & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ T_{m,1}, Role & \cdots & \cdots & \cdots & T_{m,n}, Role & D_k & CB_m & Tr_m \end{bmatrix}_{m\times(n+3)} \quad (۳-۳)$$

Term & It's Grammar Role Order based on priority — Doc NO — CB — Transition

۳-۴-تشکیل ماتریس نهایی سند-لغت

در این تحقیق از معیار وزن دهی TF-IDF برای ساخت بردار فضای اسناد استفاده شده است. برای اعمال خاصیت معنایی کلمات و اضافه کردن بار معنایی آنها در این وزن دهی، باید به نحوی تأثیر آن ها را با داخل کردنشان در این معیار نشان داد. در اینجا ما به ارائه ی دو تغییر پیشنهادی در هر دو بعد محلی و سراسری معیار TF-IDF که هر دو به صورت تجربی به دست آمده اند خواهیم پرداخت.

۳-۴-۱- تغییر اعمال شده در TF-IDF به لحاظ انتقال معنایی نظریه ی مرکزیت

در این تحقیق، تغییری را در پارامتر TF از معیار TF-IDF، و بر مبنای انتقال صورت گرفته میان هر دو جمله ی متوالی و همچنین نقش هر کلمه در هرجمله، اعمال می کنیم. پایه ی کار در این تغییرات متناسب با هر انتقال، حدس شهودی است، که بر اساس تفسیر کارکرد هر کدام از انتقال ها، به دست آمده است. در ادامه به ارائه ی تفسیر شهودی از هر انتقال و تعیین میزان اهمیت آنها به لحاظ انسجام دهی و برجسته نمایی جملاتی از متن می پردازیم.

از آنجایی که مطالعات پیرامون کاربردهای نظریه ی مرکزیت از قدمت چندانی برخوردار نمی باشد، در قسمت ۳-۴-۳ به ارائه ی یک تجربه ی آزمایشگاهی در جهت اثبات درستی حدس شهودی ارائه شده در بخش ۳-۴-۲ می پردازیم.

در قمست۳-۴-4 به نحوه ی محاسبه و ارائه ی جدول ضرایب مرتبط با هر یک از انتقال ها خواهیم پرداخت.

۳-۴-۲- حدس های شهودی درباره ی انتقال های معنایی نظریه ی مرکزیت

یک حدس شهودی در این تحقیق، بر این مبنا است، که انتقال معنایی Continue دارای بیشترین ارزش به لحاظ تعیین کنندگی میزان انسجام و برجستگی در متن می باشد. چراکه این انتقال معادل بسط یک مفهوم در قسمتی از متن است. بدین لحاظ نمی توان هیچ کدام از این جملات را کم اهمیت در نظر گرفت. انتقال

Continue در حقیقت معادل جریان یک لغت در بالاترین ارزش ممکن در سه جمله ی متوالی است. بدین جهت نیز می توان گفت که چنین لغتی باید از ارزش بالایی برخوردار باشد، چراکه در انتقال معنا در سه جمله ی متوالی دخیل بوده است. به همین دلیل و به صورت شهودی به جملاتی که انتقال معنایی Continue بین آنها رخداده است، بالاترین ضریب اثرگذاری را می دهیم.

انتقال دوم که بر اساس شهود مورد توجه قرار می گیرد، انتقال Smooth-shift می باشد. شهود مربوط به این انتقال به این صورت است که، یک لغت بین دو جمله ی متوالی در بالاترین ارزش ممکن قرار دارد. البته با ارزشترین لغت جمله ی n-2 در جمله ی n-1 نیز تکرار شده است. این بدان معنی است که معنا بین هر سه جمله در حال جریان است و هر سه با هم در حال تکمیل معنا می باشند، ولی یک تغییر معنایی بسیار ظریف میان جمله ی n-2 و جمله ی n-1 اتفاق افتاده است. از اینرو انتقال Smooth-shift دومین جایگاه را به لحاظ میزان اهمیت در انتقال معنا داراست.

انتقال معنایی Retain به لحاظ شهودی برابر با حفظ معنا میان سه جمله ی متوالی می باشد. می توان گفت که این انتقال زمانی میان دو جمله اتفاق می افتد، که تنها یک معنای واحد توسط هر دو جمله در حال بیان است. از این رو می توان حدس زد که انتساب یک وزن میانگین به کلمات جملاتی که انتقال Retain میان آنها رخداده است، بهترین انتخاب می باشد. این امر از طریق محاسبه ی یک وزن بر حسب میانگین مجموع وزن های جملات حاوی انتقال Retain قابل دستیابی می باشد.

انتقال معنایی آخر، انتقال Rough-shift می باشد. شهود مربوط به این انتقال، شامل تغییر شدید معنا میان سه جمله ی متوالی می باشد. از اینرو می توان نشان داد که در این قسمت از گفتار هیچگونه ردی از انسجام دیده نمی شود. بدین جهت می توان ادعا نمود که مفاهیم مهم این متن در قسمت از گفتار قابل بازیابی نمی باشند.

۳-۴-۳- اثبات تجربی توانایی نظریه ی مرکزیت در تشخیص انسجام و برجستگی

برای اثبات تجربی حدس های شهودی ارائه شده یک آزمایش با استفاده از اعمال نظریه ی مرکزیت در یک کاربرد زبانی- محاسباتی، با فرض قرار دادن حدس های شهودی ارائه شده در قسمت ۳-۴-۲ انجام شده است. این آزمایش تجربی که کاربرد نظریه ی مرکزیت در خلاصه سازی تک سنده بر اساس حدس های شهودی ارائه شده می باشد در فصل آینده در قسمت پیاده سازی ها مورد بحث قرار می گیرد.

۴-۴-۳- ضرایب بدست آمده برای انتقال های معنایی نظریه ی مرکزیت برای کاربرد بازیابی اطلاعات

در وزن دهی به کلمات، به ویژه در معیار مانند TF-IDF یکی از اصلی ترین رموز مؤفقیت، توجه به اهمیت محلی و سراسری هر کلمه می باشد. یک حدس شهودی استفاده شده در این تحقیق نیز داخل کردن پارامترهای محلی و سراسری در تعیین ضرایب وارده در وزن هر کلمه بر حسب انتقال معنایی صورت گرفته است. به این دلیل بدست آوردن ضرایب مربوطه در دو مرحله صورت گرفته که در ادامه به آنها اشاره می کنیم.

۱-۴-۴-۳- تعیین فرکانس مرتبط با هر کلمه در یک متن

در این مرحله با استفاده از ماتریس جمله-لغت رابطه ۳-۳ که یک ماتریس مرتب شده بر حسب میزان اهمیت هر کلمه در جمله به همراه نقش گرامری آن کلمه، و انتقال معنایی میان هر دو جمله است، ضریب مرتبط با هر شمارش فرکانس یک کلمه محاسبه می شود، تا در ادامه با تغییراتی TF آن کلمه در مجموعه ی اسناد محاسبه شود.

همانطور که پیش از این گفته شد، در پارامتر TF به ازای هر بار دیده شدن یک کلمه، یک واحد به فرکانس آن افزوده می گردد. در نهایت و پس از شمارش فرکانس کلمه در مجموعه ی اسناد و تقسیم این مقدار به کل فرکانس کلمات در مجموعه ی اسناد، پارامتر TF یک کلمه محاسبه شده است.

ابتدا باید خاطر نشان کرد که تمامی مقادیر استفاده شده در این تحقیق بر اساس تجربه به دست آمده است.

۱-۱-۴-۴-۳- محاسبه وزن هر کلمه به لحاظ انتقال معنایی رخ داده برای جمله در برگیرنده آن

ایده ی اصلی:

ایده ی اصلی در این بخش این است که، برای هر انتقال معنایی یک ضریب مشخص شود و به دو جمله ای که درگیر این انتقال می باشند اطلاق گردد. سپس در شمارش فرکانس کلمات این جملات، این ضریب در فرکانس ضرب گردد.

همینطور برای توجه بیشتر به میزان اهمیت یک کلمه در انتقال مفاهیم در سطح یک متن که یکی از نقاط ضعف بزرگ معیارهایی مانند TF-IDF می باشد، نحوه ی توزیع گرامری کلمه در سطح متن نیز در شمارش فرکانس یک کلمه در یک متن تأثیر داده می شود.

برای یک متن در تجربه های آزموده شده در این تحقیق، به منظور شمارش فرکانس یک کلمه گامهای زیر پیموده شده است:

گام یک:

در این آزمون به هر یک از انتقال های معنایی در یک متن ضریب فرکانس مطابق جدول3-1 را نسبت داده ایم. این ضریب بر اساس حدس های شهودی ارائه شده برای انتقال های معنایی و همچنین میانگین رخداد هر یک از این انتقال ها در کل مجموعه ی اسناد Medline (به عنوان مجموعه داده ی یادگیری) که در ادامه به توصیف آن خواهیم پرداخت، تعیین شده است. این ضرایب بر اساس چندین بار تکرار آزمون بازیابی اطلاعات با الگوریتم LSI در بهبود معیار دقت (آزمون F-measure) که بعداً تشریح خواهد شد، به دست آمده است.

جدول ۳-۱- ضرایب تجربی بدست آمده برای انتقال های معنایی بر اساس میانگین رخداد آنها در مجموعه داده MED

Transition	ضریب	میانگین رخداد در مجموعه ی Medline
Continue	2.3	0.135
Smooth-shift	1.4	0.17
Retain	0.6	0.28
Rough-shift	0.51	0.415

یکی از شهودهای موجود در این آزمون عدم دقت این نوع ضریب دهی به متون به هم ریخته به لحاظ توزیع مفاهیم در سطح متن می باشد. به عنوان مثال در یک متن با تعداد بالای رخداد انتقال معنایی Continue بسیاری از کلمات متن از ضریب فرکانس بالایی برخوردار می شود. بدین جهت این متن و به ویژه کلمات این متن در نهایت از مقدار TF-IDF بزرگی بهره می برند. این متن در جواب بسیاری از پرس و جو های کاربر یک کاندیدای خوب محسوب می شود. در حالیکه به هم ریختگی و توزیع نامناسب در سطح متن به لحاظ پراکندگی مفاهیم در سطح آن یکی از دلایل مهم در عدم مناسب بودن چنین کاندیدایی است.

به لحاظ آزمون های انجام شده و مشاهده متون مختلف، این ادعا ثابت شده است. حتی در متون خلاصه و یا متون خبری که یک مفهوم در حجم کمی از جملات و در سراسر متن پراکنده شده است، تعداد رخداد انتقال های معنایی نرمال بوده و هیچگاه یک انتقال معنایی به تعداد بالا در آنها رخ نمی دهد.

البته رخداد انتقال های معنایی مانند Retain و Rough-shift بدیهی به نظر می رسد. که علت آن بیهوده بودن تعداد زیادی از جملات یک متن بلند و شرکت نکردن این جملات در انتقال معنا و مفهوم آن متن می باشد.

گام دو:

در این آزمون، ابتدا به شمارش تعداد کل انتقال های معنایی و تعداد کل هر انتقال معنایی در سطح یک متن می پردازیم. سپس بر اساس میزان تکرار یک انتقال در متن و تعداد کل انتقال های متن بر اساس رابطه ی ۳-۴ به نرمال سازی آن می پردازیم. محاسبه ی این فرکانس نرمال شده از روی ماتریس جمله-لغت رابطه ی ۳-۳ به دست می آید:

وزن محاسبه شده برای هر انتقال معنایی = TrF×ITF (۳-۴)

$$TrF = \sum_{i=1}^{l} 1 \quad \textit{where l is Number of text sentences for specific transition}$$

$$ITF = \log \frac{Number\ of\ text\ sentences}{TrF}$$

بر اساس رابطه ی ۳-۴ در صورت تکرار بالای یک انتقال، ارزش آن به صورت لگاریتمی کاهش می یابد. این رابطه در حقیقت، رابطه ی اصلی محاسبه ی معیار TF-IDF می باشد.

وزن اختصاص داده به هر انتقال در متن از ضرب معیار بدست آمده از رابطه ی بالا در مقداری که به صورت تجربی و با تکرار آزمون F-measure بر روی مجموعه ی اسناد Medline بدست آمده است، محاسبه می شود. مقدار این پارامتر کمکی بر اساس جدول ۳-۲ برای هر انتقال در نظر گرفته می شود، که به صورت تجربی به دست آمده است.

جدول ۳-۲- ضرایب تجربی استفاده شده برای انتقال معنایی به دست آمده از آزمایش چندباره بر روی مجموعه داده MED

Transition	مقدار ضریب
Continue	1.8
Smooth-shift	1.3
Retain	0.8
Rough-shift	0.7

این مقدار معقول به نظر می رسد، چراکه ارزش انتقال هایی مانند Retain و Rough-shift که به صورت طبیعی بالا و بی ارزش هستند، کاهش می یابد. همچنین در متونی که به لحاظ توزیع مفهوم، پراکندگی بالایی دارند، ارزش یک انتقال معنایی از هر نوعی با تکرار بالا، کاهش می یابد. این مسئله خود به کاهش ارزش آن متن به لحاظ اهمیت برای کاندیدا شدن در جواب پرس و جو ها می انجامد.

نتایج بدست آمده از تجارب آزمایشگاهی ما، نشان می دهد که، این نوع وزن دهی، جواب های بهتری را در اختیار می گذارد.

حال مقدار به دست آمده برای هر انتقال در عدد یک به ازای دیده شدن یکبار از یک کلمه در یک جمله ضرب می شود. شایان ذکر است که برای هر جمله، انتقال معنایی با ارزش بالاتر به لحاظ مقدار به دست آمده برای آن انتقال در متن، در نظر گرفته می شود. به عنوان مثال اگر جمله ی n ام با جمله ی n-1 ام دارای انتقالی با ارزش 1و با جمله ی n+1 ام دارای انتقالی با ارزش 2 باشد، انتقال دوم برای جمله ی n ام در نظر گرفته می شود.

حاصل این گام تغییر ماتریس جمله-لغت که یک ستون جدید به آن اضافه شده است. این ستون جدید حاوی ضریب وزن دهی مبتنی بر انتقال معنایی برای کلمات یک جمله می باشد. این ماتریس در رابطه ی ۳-۵ مشاهده می شود.

Sentence-Term Matrix: (۳-۵)

Term & It's Grammar Role

Order based on priority

Doc NO CB Transition Transition Weight

$$\text{Sentence}\begin{bmatrix} T_{1,1}, Role & \cdots & \cdots & \cdots & T_{1,n}, Role & D_1 & CB_1 & Tr_1 & TW_1 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & T_{i,j}, Role & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ T_{m,1}, Role & \cdots & \cdots & \cdots & T_{m,n}, Role & D_k & CB_m & Tr_m & TW_m \end{bmatrix}_{m\times(n+4)}$$

۲-۱-۴-۴-۳- محاسبه وزن محلی هر کلمه به لحاظ نقش گرامری آن

حال برای توجه بیشتر به معنا در سطح یک متن که یکی از نقاط ضعف بسیاری از روش های کنونی وزن دهی به کلمات می باشد، عملیات این بخش را انجام خواهیم داد. یکی از مزایای توجه به نقش گرامری کلمات در محاسبه ی وزن مربوط به آنها پر شدن خلأ مربوط به انتقال معنایی است. گاهاً کلمه ای در یک نقش گرامری مهم مانند فاعل به دفعات تکرار شده است، اما در جملاتی که میان آنها یک انتقال معنایی با ارزش صورت گرفته، مشاهده نشده است. توجه به نقش گرامری کلمات و دخیل کردن آن در وزن دهی می تواند این خلأ ذاتی در روش وزن دهی پیشنهادی را تا حد زیادی جبران کند. در این گام کلیه ی نقش های گرامری یک کلمه را شمارش می کنیم. این نقش ها در ماتریس جمله-لغت برای هر کلمه مشخص شده است. به عنوان مثال تعداد دفعاتی که یک کلمه در یک متن به عنوان فاعل انتخاب شده است را می شماریم. به همین ترتیب برای نقش های دیگر نیز تعداد رخدادهای برای هر کلمه در متن را شمارش می کنیم. حال ماتریس موقتی سند-کلمه/نقش را که سطرهای آن کلمه-نقش و ستون های آن اسناد و هر خانه ی آن مقدار تکرار یک کلمه-نقش می باشد را تشکیل می دهیم. این ماتریس در رابطه ی ۳-۶ مشاهده می شود. در این ماتریس m تعداد کل اسناد و n حداکثر تعداد کلمه-نقش در یک سند می باشد.

Document-Term/Role: (۳-۶)

Term/Role/Frequency

$$\text{Documents}\begin{bmatrix} Term_{1,1}, Role, Freq & \cdots & \cdots & Term_{1,n}, Role, Freq \\ \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ Term_{m,1}, Role, Freq & \cdots & \cdots & Term_{m,n}, Role, Freq \end{bmatrix}_{m\times n}$$

سپس برای هر کلمه در یک متن تعداد کل تکرار ها در همه ی نقش های گرامری و تعداد کل تکرار آن کلمه در یک نقش خاص مانند فاعل از روی ماتریس سند-کلمه/نقش شمارش شده و بر اساس رابطه ی ۳-۷ مقدار ضریب نقش مورد نظر برای کلمه محاسبه می شود.

وزن محاسبه شده برای هر نقش در یک متن = TRF/TF (۳-۷)

Where TRF = $\sum Freq,\ for\ specific\ role\ e.g.\ subject$

TF is total ferequence for term in a text

علت نرمال شدن این مقدار در کل متن، ویژگی های زبانی است. به عنوان مثال مقالات که معمولاً با جملات مجهول نوشته می شوند از تعداد کم فاعل در کل متن بهره می برند. به صورت شهودی می توان گفت که در این نوع از اسناد، مفعول از جایگاه مهم تری برخوردار است.

سپس مقدار به دست آمده در ضریبی که برای هر نقش به صورت تجربی و بر اساس آزمایش چندباره ی F-measure مطابق جدول ۳-۳ به دست می آید، ضرب می شود.

جدول ۳-۳- ضرایب استفاده شده برای نقش های گرامری، به دست آمده از آزمایش چندباره بر روی مجموعه داده MED

نقش گرامری	ضریب مبتنی بر آزمایش برای نقش گرامری
فاعل	۳,۲۴۵
صفت مرتبط با فاعل	۲,۹۸۷
مفعول غیر مستقیم	۲,۴۵۱
صفت مرتبط با مفعول غیر مستقیم	۲,۳۴۲
مفعول مستقیم	۲,۲۳۲
صفت مرتبط با مفعول مستقیم	۲
قید مکان	۰,۸۷۶
صفت مرتبط با قید مکان	۰,۷۵۶

حال میانگین مقدارهای به دست آمده برای نقش های مختلف یک کلمه در متن را محاسبه کرده و در فرکانس به دست آمده برای هر کلمه از نظر انتقال معنایی که در ماتریس جمله-لغت آمده است، بر اساس فرمول ۳-۸ ضرب می کنیم.

$$\text{فرکانس یک کلمه در یک جمله} = \frac{\sum_{i=1}^{j:number\ of\ variuos\ role\ for\ specific\ term} TrF/TF}{\sum_{i=1}^{j:number\ of\ variuos\ role\ for\ specific\ term} 1} \times \text{TW} \qquad (۳-۸)$$

تاکنون مقدار فرکانس هر کلمه در هر جمله از هر متن مشخص شده است. این مقدار را به خانه ی متناظر در ماتریس جمله-لغت اضافه می کنیم. ماتریس جدید در رابطه ی ۳-۹ مشاهده می شود.

Sentence-Term Matrix: (۳-۹)

Term & It's Grammar Role Order based on priority &It's weight — Doc NO — CB — Transition — Transition Weight

Sentence

$$\begin{bmatrix} T_{1,1}, Role, W_{1,1} & \cdots & \cdots & \cdots & T_{1,n}, Role, W_{1,n} & D_1 & CB_1 & Tr_1 & TW_1 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & T_{i,j}, Role, W_{i,j} & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ T_{m,1}, Role, W_{m,1} & \cdots & \cdots & \cdots & T_{m,n}, Role, W_{m,n} & D_k & CB_m & Tr_m & TW_m \end{bmatrix}_{m\times(n+4)}$$

در ادامه به شرح راه حل به کار رفته جهت دخیل کردن اهمیت سراسری یک کلمه به لحاظ حمل بار معنایی در مجموعه ای از اسناد و ترکیب آن با مقادیر فرکانس به دست آمده تاکنون می پردازیم.

۵-۴-۳- تعیین فرکانس مرتبط با هر کلمه در مجموعه ای از اسناد

تکرار یک کلمه در یک متن، به دفعات در یک نقش گرامری خاص و مهم، مانند فاعل، واقعاً نشان دهنده ی فعال بودن آن کلمه در انتقال مفاهیم مربوط به آن متن می باشد. شاهد این امر [GAG05] و [AMI04] و [SID04] هستند که بر اساس نقش گرامری کلمات دست به انتخاب جملات مهم متن و خلاصه ی آن زده اند.

اما در مورد نقش گرامری یک کلمه در مجموعه ای از اسناد، نمی توان با این قاطعیت نظر داد. چراکه به عنوان مثال نمی توان گفت یک کلمه که در بسیاری از اسناد در نقش قید به کار رفته، تأثیر پایینی در انتقال مفاهیم در مجموعه ی اسناد داراست.

در این قسمت به مانند گام سوم بخش گذشته به شمارش تکرار یک کلمه در یک نقش خاص در مجموعه ی اسناد از روی ماتریس جمله-لغت اقدام می کنیم. همچنین تعداد کل رخداد نقش گرامری مورد نظر در مجموعه ی اسناد را می شماریم. حال برای این کلمه در نقش گرامری مورد نظر، مقداری را بر اساس رابطه ی ۳-۱۰ محاسبه می کنیم.

فرکانس کلمه بر اساس نقش گرامری آن در کل مجموعه ی اسناد = (۳-۱۰)

$$\mathrm{TRW} = \frac{\sum_{i=1}^{n:\,number\ of\ specific\ term\ for\ specific\ role\ e.g.subject\ in\ hole\ of\ dataset} 1}{\sum_{i=1}^{n:number\ of\ specific\ term\ in\ hole\ of\ dataset} 1} \times log \frac{\sum_{i=1}^{n:all\ of\ dataset\ terms} 1}{\sum_{i=1}^{n:all\ of\ specific\ role\ in\ dataset\ e.g.subject} 1}$$

سپس به مانند گام سوم از بخش گذشته، میانگین مقادیر به دست آمده برای هر کلمه در نقش های مختلف را محاسبه کرده و در مقدار فرکانس $W_{i,j}$ این کلمه در ماتریس جمله-لغت مطابق رابطه ۳-۱۱ ضرب می کنیم. سپس این مقدار را جایگزین مقدار $W_{i,j}$ در ماتریس جمله-لغت می کنیم.

$$\text{فرکانس کلمه در کل مجموعه ی اسناد} = \frac{\sum_{k=1}^{l:number\ of\ variuos\ role\ for\ specific\ term} TRW}{\sum_{k=1}^{l:number\ of\ variuos\ role\ for\ specific\ term} 1} \times W_{i,j} \quad (۳-۱۱)$$

۶-۴-۳- محاسبه ی TF-IDF

در محاسبه ی مقدار TF به ازای هر بار تکرار یک کلمه، یک واحد به فرکانس شمارش اضافه می کنیم. اما در این تحقیق دیدیم که ماتریس جمله-لغت شامل کلیه ی کلمات است که در بسیاری از موارد به ازای یک بار دیده شدن آنها، مقدار متفاوتی از یک (بیشتر و یا کمتر) رخ می دهد. پارامتر TF شامل مقدار فرکانس یک کلمه تقسیم بر کل فرکانس کلمات در یک متن می باشد. حال با جمع فرکانس یک کلمه در یک متن از ماتریس جمله-لغت بر حسب مقدار $W_{i,j}$ آن و جمع کل فرکانس کلمات از ماتریس جمله-لغت بر حسب اضافه شدن یک واحد در ازای دیدن یکبار کلمه، این مقدار را تشکیل می دهیم. رابطه ی محاسبه کننده این مقدار در ۳-۱۲ دیده می شود.

$$\text{TF} = \frac{\sum^{In\ a\ text} W_{Term}}{\sum^{In\ a\ text} 1_{for\ each\ Term}} \quad (۳-۱۲)$$

برای محاسبه ی فرکانس کل کلمات موجود در مجموعه ی اسناد و یا یک متن نیز از ماتریس جمله-لغت استفاده می کنیم و هر کلمه به ازای یکبار دیده شدن، یک واحد به این مقدار اضافه می کند. علت این کار این است که، در صورت استفاده از مقدار $W_{i,j}$ برای محاسبه ی مقدار کل فرکانس کلمات، تأثیر تغییرات داده شده و گام های پیموده شده در بالا برای محاسبه ی TF، کاهش خواهد یافت، که این موضوع، مطلوب نمی باشد.

همانطور که قبلاً دیدیم، پارامتر TFجزء وزن های محلی به حساب می آید. هرچند که در این تحقیق سعی شد تا با دخالت نقش گرامری کلمات در سراسر مجموعه، تأثیر وزن سراسری نیز در این پارامتر لحاظ شود.

در نهایت مقدار معیار TF-IDF بر اساس رابطه ی ۳-۱۳ و با استفاده از مقدار TF که برای هر کلمه از رابطه ی ۳-۱۲ بدست آمد، محاسبه می شود. در این رابطه |D| تعداد کل اسناد و d تعداد اسنادی است که کلمه ی مورد نظر در آن دیده می شود.

$$TF_i \times \log \frac{|D|}{|\{d:TF_i \in d\}|} \qquad (۳-۱۳)$$

خروجی این مرحله، که گام نهایی در تکمیل معیار وزن دهی معنایی می باشد ماتریس کلمه- سند TF-IDF می باشد. این ماتریس مورد استفاده ی الگوریتم های بازیابی اطلاعات مانند LSI قرار می گیرد، که در رابطه ی ۳-۱۴ مشاهده می شود. در این ماتریس n تعداد کل اسناد و m تعداد تمام کلمات موجود در همه ی اسناد هستند. برای کلماتی که در یک متن موجود نیستند مقدار صفر را در نظر می گیریم.

Term-Document Matrix: (۳-۱۴)

$$\text{Documents} \begin{bmatrix} TF-IDF_{1,1} & \cdots & \cdots & \cdots & TF-IDF_{1,m} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & TF-IDF_{i,j} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ TF-IDF_{n,1} & \cdots & \cdots & \cdots & TF-IDF_{n,m} \end{bmatrix}_{n\times m} \quad \text{Terms}$$

۵-۳- خلاصه

در این فصل به ارائه روش جدید برای وزن دهی به کلمات که هر دو ویژگی معنایی و گرامری زبان را پوشش می دهد پرداختیم. این روش چالش های مهم عدم دقت و کارایی، وابسته بودن به داتش اولیه، حجم پردازش اولیه بالا به دلیل وابستگی به مجموعه دانش اولیه و تغییر داده ی اولیه را حل نموده است. هرچند که روش پیشنهادی ارائه شده در این فصل نسبت به روش های آماری وزن دهی به کلمات از مرتبه زمانی اجرایی بالاتری برخوردار است. این مرتبه زمانی بالاتر به علت پردازش های زبانی بیشتری است که به دلیل اعمال دید معنایی و گرامری در وزن دهی لازم می باشند. البته به درستی می توان از این افزایش مرتبه زمانی چشم پوشی کرد، چرا که در بهترین حالت در کاربردهای بازیابی اطلاعات مرتبه زمانی از درجه O(m logn) است که در آن m تعداد اسناد و n تعداد کلمات می باشند. همانطور که می دانیم در دنیای واقعی تعداد اسناد بسیار زیاد می باشند.

الگوریتم پیشنهادی شامل توجه به دو ویژگی از زبان یعنی معنا و گرامر می باشد. به این منظور یک معیار خوب محلی به نام TF انتخاب شده است تا با تغییر آن به روش جدیدی دست پیدا کنیم. علت انتخاب تنها یک معیار محلی، عدم توانایی تئوری پایه معنایی روش پیشنهادی یعنی نظریه مرکزیت در تشخیص انسجام و برجستگی سراسری می باشد.

الگوریتم پیشنهادی با وزن دهی به انتقال های معنایی رخ داده میان جملات به هر کدام از کلمات این جملات وزنی تخصیص می دهد. سپس بر اساس نقش گرامری کلمات و نحوه توزیع این نقش های گرامری

وزن جدیدی را در وزن اختصاص داده شده قبلی ضرب می کنیم. سپس برای پر کردن خلأ ناشی از عدم توجه به اهمیت سراسری کلمات، وزنی را بر اساس نقش گرامری آنها در سند و توزیع این نقش گرامری در مجموعه کل اسناد، در وزن قبلی ضرب می کنیم.

در محاسبه مقدار TF هر کلمه به ازای دیده شدن هر کلمه یکبار به فرکانس شمارش آن اضافه می شود. سپس برای نرمال کردن این مقدار، آنرا بر تعداد کل کلمات سند تقسیم می کنیم. اما در روش پیشنهادی، به ازای دیده شدن هر کلمه مقدار وزن جدید آن که ممکن است از یک بیشتر و یا کمتر باشد، به فرکانس شمارش شده آن تاکنون اضافه می شود. به هنگام تقسیم این مقدار، آنرا بر همان تعداد کلمات متن تقسیم می کنیم، تا از تأثیر معیار وزن دهی جدید کاسته نشود.

علاوه بر اضافه شدن زمان اجرا، تنها چالش بزرگ دیگر پیش رو روش پیشنهادی وابسته بودن آن به زبان می باشد. زیرا این روش از نظریه مرکزیت که یک تئوری زبانی وابسته به زبان مبدأ می باشد، استفاده می کند.

فصل چهارم:

پیاده سازی و ارزیابی

۴– پیاده سازی و ارزیابی

۴–۱– مقدمه

ساختار کلی این فصل به سه قسمت اصلی تقسیم شده است.

در بخش اول از این فصل (۴-۲) ابتدا توضیحی کوتاه در مورد پیاده سازی های لازم برای اجرای روش وزن دهی پیشنهادی ارائه خواهد شد. این پیاده سازی ها تنها شامل پیش پردازش های لازم برای آماده سازی مجموعه لغات جهت انجام وزن دهی بر روی آنها می شود. مراحل پیش پردازشی لازم، هم در فصل مرور ادبیات و هم در فصل روش پیشنهادی مورد بررسی و معرفی قرار گرفتند.

در بخش دوم از این فصل (۴-۳) به ارائه تجربه آزمایشگاهی که بر اساس آن چهار حدس شهودی ارائه شده در فصل روش پیشنهادی، حول چهار انتقال معنایی نظریه مرکزیت، ارائه شد، به صورت تجربی نیز اثبات می شوند.

در بخش سوم از این فصل (۴-۴) نیز، به ارائه آزمایش های انجام گرفته، جهت ارزیابی روش پیشنهادی و مقایسه آن با روش های مشهور وزن دهی، پرداخته شده است.

۴–۲– پیاده سازی

همانطور که گفته شد، برای اجرای الگوریتم پیشنهادی بر روی مجموعه ای از داده های که خروجی نهایی آن ماتریس وزن دهی شده ی سند-لغت می باشد، احتیاج به پیمودن گام هایی در پیش پردازش مجموعه ی داده های به منظور آماده سازی آن برای اجرای الگوریتم مورد نظر می باشد. در این قسمت، اشاره ای اجمالی به گام هایی که برای انجام این عملیات باید پیموده شوند خواهیم داشت.

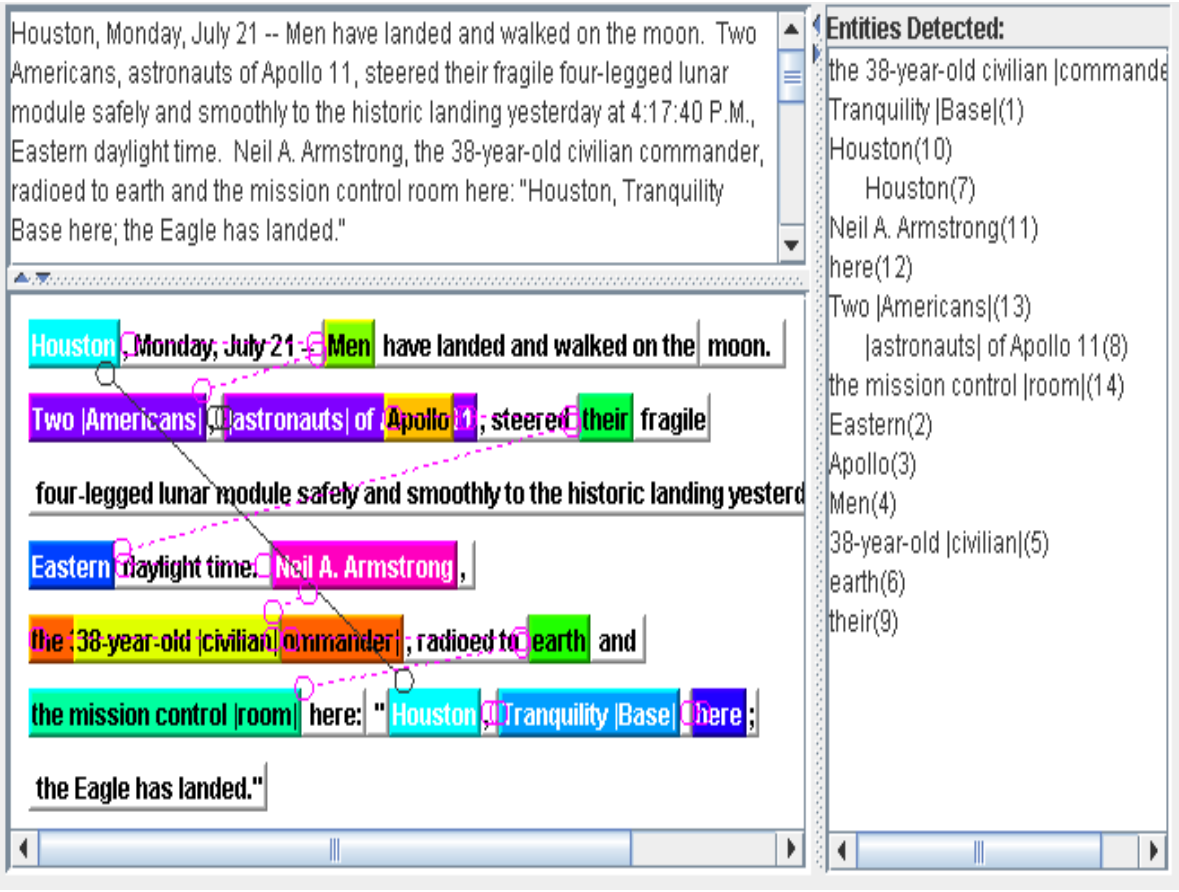
۴–۲–۱– تشخیص زنجیره های مرجعیتی

برای انجام ماشینی این پردازش متنی ابزارهای آماده ای که به صورت آکادمیک تولید شده است، نیز وجود دارد. در این تحقیق از ابزار آزمایشگاه محاسبات علوم شناختی[1] دانشگاه Illinois At Urbana Champaign استفاده شده است که با دقت خوبی عملیات جانشینی آنافورا را انجام می دهد. هم اکنون تنها یک آمار مبنی بر میانگین دقت ۷۵٪ برای روش های جانشینی وجود دارد [ELA05]. از آنجا که تاکنون ارزیابی دقیقی در مورد

[1] - Cognitive Computation

دقت روش های جانشینی ارائه نشده است، انتخاب ابزار دانشگاه Illinois که از یک روش یادگیری ماشین استفاده کرده است، به علت دقت ارائه شده در [BEN08] است که دارای B-cube F 80.8% می باشد.

ورودی این ابزار یک متن می باشد و یکی از خروجی های آن عناصری وابسته به همراه تگ مربوط به مرجع آن ها است. ما توانستیم با اعمال تغییراتی در کد ابزار متن باز دانشگاه Illinoise خروجی این نرم افزار را به یک متن تصحیح شده با جایگذاری مراجع هر یک از عناصر وابسته تبدیل کنیم. یک نمونه ی گرافیکی از خروجی این ابزار در شکل ۴-۱ دیده می شود.



شکل ۴-۱- نمای گرافیکی از خروجی ابزار تشخیص زنجیره های مرجعیتی

۲-۲-۴- برچسب زنی معنایی نقش کلمات

در این مرحله از ابزار برچسب زنی معنایی نقش کلمات گروه محاسبات علوم شناختی دانشگاه Illinois At Urbana Champaign استفاده کردیم. این ابزار از پارسر Charniak [CHA97] برای برچسب زنی به عناصر جمله استفاده می کند. الگوریتم مورد استفاده در این ابزار از دسته ی دوم الگوریتم های برچسب زنی معنایی محسوب می شود.

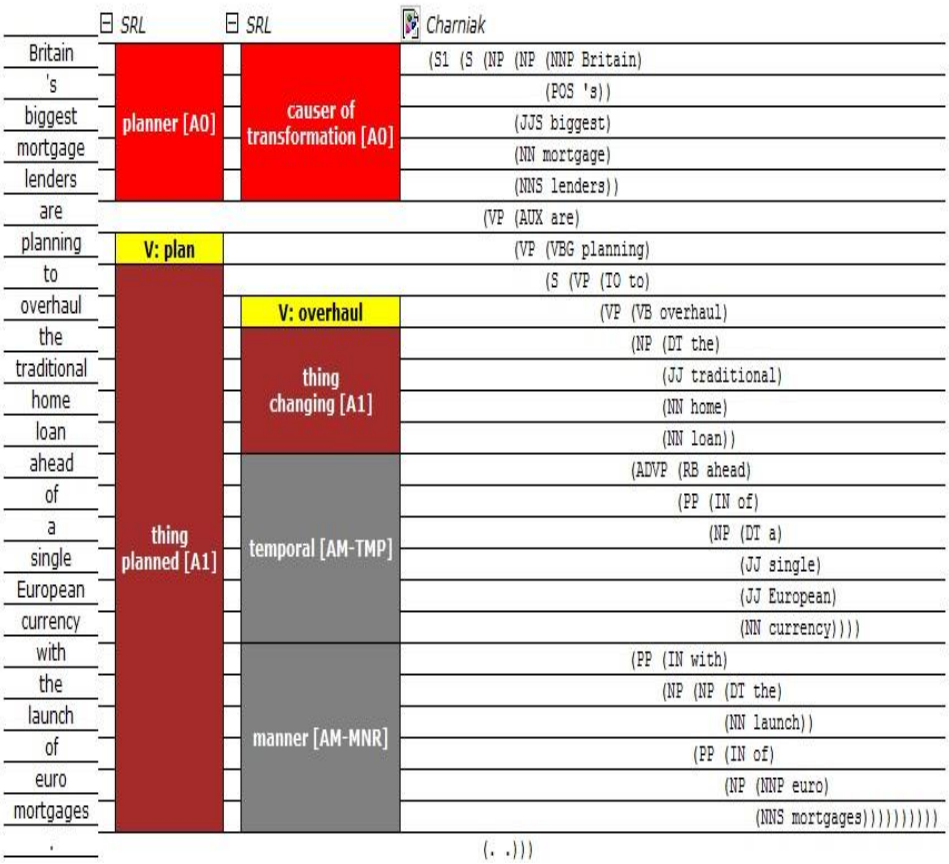
خروجی این ابزار به صورتی ماتریسی است که سطرهای آن کلمات جمله و ستون های آن سطوح شکسته شدن جمله بر اساس افعال موجود در جمله است. در شکل ۴-۲ یک نمای گرافیکی از چنین ماتریسی مشاهده می شود.

همانطور که مشاهده می شود، به عنوان مثال، مفعول خود می تواند، یک جمله باشد. در حالیکه احتیاج است که تنها یک اسم به عنوان مفعول باز گردانده شود. در نتیجه در این مرحله، باید حداقل یک گروه اسمی به عنوان مفعول برگردانده شود، تا پس از اعمال برچسب زنی نحوی یکی از این اسامی به عنوان مفعول نهایی انتخاب شود. برای مشخص کردن فاعل اصلی جمله، بزرگترین سطح پردازشی برچسب زنی معنایی نقش کلمات که سطحی است که تمامی جمله را در بر می گیرد، به عنوان سطح اول انتخاب می کنیم. اگر مانند، مثال فوق دو سطح که تمامی جمله را در بر می گیرند، موجود باشد، سطحی که بزرگترین نقش فاعلی یا مفعولی و کوچکترین نقش قیدی را دارد، به عنوان سطح اول در نظر گرفته می شود. سطح های بعدی نیز بر همین اساس و بر اساس حداکثر تحت پوشش قرار دادن جمله تعیین می گردد.

Britain's biggest mortgage lenders are planning to overhaul the traditional home loan ahead of a single European currency with the launch of euro mortgages.

Click For General Explanation of Argument Labels

Output:



شکل ۴-۲- نمای گرافیکی از خروجی ابزار برچسب زنی معنایی نقش کلمات

پس از مشخص شدن یک نقش مانند مفعول، اگر خود مفعول یک جمله باشد مانند شکل ۴-۲، مفعول به سطح های بعدی خود شکسته می شود تا در نهایت به یک کلمه و یا یک گروه اسمی به عنوان مفعول برسیم.

۴-۲-۳- برچسب زنی نحوی لغات

در این تحقیق، از ابزار برچسب زنی POS گروه تحقیقاتی محاسبات علوم شناختی دانشگاه Illinois at Urbana Champaign برای عمل برچسب زنی نحوی استفاده شده است. این ابزار از دقت خوبی برخوردار می باشد.

۴-۲-۴- حذف کلمات زائد

این عملیات از روی دیکشنری کلمات زائد انجام می پذیرد. علت استفاده نکردن از محصولات آماده در این بخش، پیچیدگی خروجی آنها و نوشتن کدهای اضافه برای پارس این خروجی ها و همچنین سادگی پیاده سازی شخصی این الگوریتم ها می باشد. هرچند که ابزارهای دقیق و کارایی مانند ابزارهای دانشگاه Illinois at Urbana Champaign موجود می باشد.

۴-۲-۵- ریشه یابی

عملیات این گام بر روی ماتریس های تولید شده در گام های پیشین صورت می پذیرد. علت استفاده نکردن از محصولات آماده در این بخش، پیچیدگی خروجی آنها و نوشتن کدهای اضافه برای پارس این خروجی ها و همچنین سادگی پیاده سازی شخصی این الگوریتم ها می باشد. هرچند که ابزارهای دقیق و کارایی مانند ابزارهای دانشگاه Illinois at Urbana Champaign موجود می باشد.

۴-۲-۶- تشخیص انتقال های معنایی میان جملات

به صورت خودکار و بر مبنای ماتریس تولید شده جمله-لغت که در آن پارامتر CB هر جمله نیز مشخص شده است و بر اساس تعریف ارائه شده برای قوانین انتقال معنایی انجام می پذیرد.

۴-۳- آزمایش درستی حدس های شهودی درباره انتقال های معنایی

در این بخش به ارائه یک آزمایش تجربی برای اثبات حدس های شهودی درباره انتقال های معنایی نظریه مرکزیت که در قسمت ۳-۴-۲ مشاهده شد، می پردازیم.

۴-۳-۱- خلاصه سازی بر مبنای نظریه ی مرکزیت

خلاصه‌سازی خودکار متون یکی از ابزارهای مهم و در عصر افزایش انفجارگونه ی اطلاعات می‌باشد. طبق تعریف [RAD02] خلاصه به یک متن تولیدشده از یک یا چند متن دیگر اطلاق می‌شود که دربردارنده مفاهیم مهم متن یا متون مبدأ می‌باشد. این متن تولیدشده نباید از نصف متن یا متون مبدأ بزرگتر باشد. همین تفسیر ساده دربردارنده ویژگی‌های اصلی یک خلاصه می‌باشد: خلاصه یک یا چند متن، دربرداشتن اطلاعات مهم متون مبدأ و کوتاه بودن.

تحقیقات برای کشف دانش مهم و برجسته در داخل یک متن حول موضوع خلاصه‌سازی تک‌سنده می‌باشد [DAS07]. این تحقیقات پیرامون دو دسته تولید خلاصه‌های چکیده‌ای[1] و استخراجی[2] بسط پیدا کرده‌اند. خلاصه استخراجی به معنی برگرداندن تعدادی از جملات متن به عنوان قسمت‌های مهم و خلاصه چکیده‌ای به معنای بیان دانش درونی یک متن به هر شکل ممکن می‌باشد [DAS07].

در این آزمایش به ارائه یک روش استخراجی خلاصه‌سازی برای یک متن، که تلفیقی از یک نظریه زبانشناسی به نام نظریه مرکزیت و برخی پارامترها و فرمول‌های آماری است می‌پردازیم. این تلفیق به منظورکاهش آثار مشکلات روش‌های کنونی خلاصه‌های استخراجی ارائه شده است. بیشتربودن طول جملات استخراجی از میانگین طول جملات متن، پراکندگی اطلاعات در سطح متن، تداخل اطلاعات جملات استخراج شده و عدم انسجام در خلاصه تولیدشده، از جمله این مشکلات می‌باشد. پارامترهای آماری به منظور حل مشکل اول و نظریه مرکزیت ایده اصلی برای حل باقی مشکلات می‌باشد. در بخش دوم این آزمایش مروری بر ادبیات خلاصه‌سازی تک‌سنده استخراجی و نظریه مرکزیت خواهیم داشت. در بخش سوم روش پیشنهادی و تعدادی از زیرالگوریتم‌های لازم برای پیش-پردازش مورد بحث قرار می‌گیرند. در بخش چهارم نتایج مقایسه و ارزیابی روش پیشنهادی با تعدادی از سیستم-های موجود ارائه شده‌است.

۱-۱-۳-۴- خلاصه‌سازی تک‌سنده استخراجی

روش‌های زیادی برای خلاصه‌سازی یک متن ارائه شده‌است که به دلیل کثرت کارهای انجام شده به مهمترین آن-ها اشاره می‌کنیم. ادمونسون با ترکیب چهار پارامتر فرکانس کلمه، موقعیت جمله، توجه به کلمات نشانه‌ای مانند *این مهم است که، از آنجاکه و ....* و توجه به اسکلت جمله مانند کوتاه‌بودن و یا سربخش‌بودن آن، روش جدیدی برای خلاصه‌سازی ارائه کرد [DAS07]. در یک پژوهش دیگر [GON01] از روش LSA به عنوان یک روش خوشه بندی که از یک معیار وزن‌دهی لگاریتمی استفاده می‌کند، برای خلاصه‌سازی بهره برده شده است. در تحقیق دیگری با استفاده از یک شبکه عصبی برای پردازش داده‌های DUC2001-02 نتیجه‌گیری شده است که اولین جمله هر متن خبری، مهمترین جمله آن به شمار می‌رود [DAS07]. در [HOF96] روشی مبتنی بر استفاده از زنجیره لغوی[3] ارائه شده‌است. همچنین [HOF96] روشی مبتنی بر نظریه مرکزیت برای خلاصه‌سازی ارائه کرده‌است. در این روش با محاسبه CB هر جمله و اندازه‌گیری CBهای مشابه در متن، جملاتی که پرتکرار-ترین CBها را در خود داشته‌باشند، به عنوان خلاصه برگردانده می‌شوند.

---

[1]- Abstractive
[2]- Extractive
[3]- Lexical chain

۲-۱-۳-۴- انتخاب جملات مهم بر مبنای نظریه ی مرکزیت

در این تحقیق برای انتخاب جملات مهم گام‌هایی پیموده شده‌است که به ترتیب به آن‌ها اشاره خواهد شد.

۱-۲-۱-۳-۴- وزن‌دهی به جملات بر اساس انتقال انجام‌شده میان دو جمله متوالی

در این تحقیق بر اساس یک حدس اولیه که مبتنی بر ساختار انتقال‌های معنایی می‌باشد، برخی از جملات بر حسب انتقال معنایی رخ داده میان آن‌ها به عنوان جملات مهم انتخاب شده‌اند. انتقال‌های معنایی در نظریه مرکزیت، طبق آنچه نشان داده‌شد، در حقیقت ارتباط میان کلمات دو جمله متوالی می‌باشند. بر این اساس هر بار یکی از انتقال‌های معنایی مورد توجه قرار گرفته‌است و به‌ترتیب هر کدام از این انتقال‌ها مورد ارزیابی قرارگرفته‌اند. برای هر کدام از انتقال‌ها میانگین رخداد آن‌ها میان جملات یک متن در کل پیکره مورد ارزیابی به دست‌آمده‌است که مطابق جدول شماره 4-۱ قابل مشاهده است. بر این اساس نوع امتیازدهی به این انتقال‌ها برای انتخاب جملات مهم مطابق رابطه۴-۱ می‌باشد[کام۸۹].

$$TF = \left(\frac{f_{transition,i}}{\sum_k f_{transition,k}}\right) Log_2 \left(\frac{|D|}{|\{d:transition \in d\}|}\right) \qquad (۴-۱)$$

مقدار به دست‌آمده برای هر انتقال، یک مقدار عددی ثابت است که از محاسبه مقادیر گفته‌شده بر روی DUC2001 به عنوان مجموعه داده یادگیری به دست آمده‌است. بررسی‌ها و نتایج به‌دست‌آمده، دقت مناسب این ضریب محاسبه‌شده را نشان می‌دهند.

جدول ۴-۱- میانگین رخداد هر انتقال معنایی در DUC2001 [کام۸۹]

Transition	Occur Average in DUC2001
Continue	0.15
Retain	0.24
Smooth-shift	0.19
Rough-shift	0.42

۱-۱-۲-۱-۳-۴- وزن‌دهی به انتقال‌ها بر اساس فرکانس تکرار در متن

بر اساس یک شهود می‌توان حدس زد که میزان تکرار بالای یک انتقال میان جملات یک متن از اعتبار آن انتقال می‌کاهد. به همین دلیل با محاسبه فرکانس نسبی تکرار هر انتقال در هر متن و نرمال‌کردن آن این مشکل را برطرف خواهدشد. رابطه ۴-۲ فرمول محاسبه این فرکانس نسبی نرمال‌شده را نشان می‌دهد [کام۸۹].

$$TW_n = \beta * (MT_n - MA) / VA * TF \qquad (۴-۲)$$

Where $VA := {}^{1}/_{4} \sum_{i=1}^{4} (MT_i - MA)^2$

And $MT_n := \frac{\sum_{i=1}^{number\ of\ this\ transition\ n} 1}{\sum_{k=1}^{number\ of\ transitions} 1}$

And $MA := {}^{1}/_{4} \sum_{j=1}^{4} \frac{\sum_{i=1}^{number\ of\ this\ transition\ j} 1}{\sum_{k=1}^{number\ of\ transitions} 1}$

بر اساس آزمایش‌های صورت‌گرفته مقدار ضریب β برای هر انتقال مطابق جدول شماره ۴-۲ به‌دست آمده‌است.

جدول ۴-۲- مقدار تجربی ضریب β برای انتقال های معنایی [کام۸۹]

Transition	β Factor
Continue	2.6
Retain	1
Smooth-shift	1.8
Rough-shift	0.85

۴-۳-۱-۲-۲- محاسبه فرکانس نسبی کلمه–جمله و نرمال کردن آن

حجم جملات استخراج‌شده بر حسب انتقال‌ها در برخی از موارد بیشتر از حجم لازم برای خلاصه می‌باشد. برای کاهش حجم گام‌های زیر را مورد استفاده قرار داده خواهد شد.

۴-۳-۱-۲-۱- وزن‌دهی به کلمات

برای هر کلمه در جمله بر اساس رابطه ۴-۳ مقدار TF-ISF محاسبه می‌گردد [کام۸۹].

$$(TF * ISF)_{j,i} = \left(\frac{tf_{j,i}}{\sum_k tf_{j,k}}\right) Log_2 \left(\frac{|D|}{|\{d: t_i \in d\}|}\right) \qquad (۴-۳)$$

در این رابطه $tf_{j,i}$ مقدار تکرار یک کلمه در یک جمله، $|D|$ تعداد کل جملات یک متن و $|\{d: t_i \in d\}|$ تعداد کل جملاتی است که کلمه مورد نظردر آن‌ها دیده شده‌است. حال با استفاده از رابطه ۴-۴ میانگین وزن کلمات یک جمله را محاسبه و آنرا نرمال می‌کنیم [کام۸۹].

$$SMT_j := \frac{\sum_{i=1}^{n} NTFPS_{j,i}}{\sum_{i=1}^{n} NTFPS_{j,i}>0} \quad (۴-۴)$$

Where $\text{NTFPS}_{j,i} := \left(\left(\frac{tf_{j,i}}{\sum_k tf_{j,k}}\right) - MTPS\right) / VPS$

And $MTPS_j := \frac{\sum_{i=1}^{n}\left(\frac{tf_{j,i}}{\sum_k tf_{j,k}}\right)}{\sum_{i=1}^{n} tf_{j,i}>0}$

And $VPS_j := \frac{\sum_{i=1}^{n}\left(\left(\frac{tf_{j,i}}{\sum_k tf_{j,k}}\right) - \sum_{i=1}^{n}\left(\frac{tf_{j,i}}{\sum_k tf_{j,k}}\right)\right)^2}{\sum_{i=1}^{n} tf_{j,i}>0}$

حال مقدار وزن به‌دست‌آمده برای هر جمله را در وزن به‌دست‌آمده از رابطه ۴-۴ ضرب می کنیم.

۳-۱-۳-۴- ارزیابی کارایی

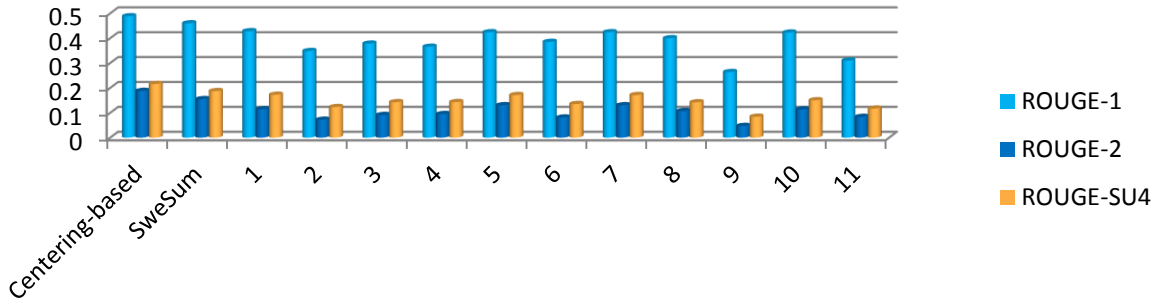
۱-۳-۱-۳-۴- مجموعه داده‌ها

همانطور که پیشتر گفته شد در ارزیابی و پیاده سازی این الگوریتم از مجموعه داده‌هایی با عنوان DUC2001 استفاده شده است. ساختار کلی این مجموعه ها به صورت تعدادی کلاستر است، که در هر کلاستر چندین متن خبری با یک موضوع از خبرگزاری های معتبر موجود می باشد. به این جهت برای هر متن و هر کلاستر خلاصه های انسانی چکیده ای در چند حجم مختلف موجود می باشد. در این مقاله تمرکز بر روی خلاصه های تک سنده با حجم ۱۰۰ کلمه‌ای می باشد. DUC2001 شامل 600 سند می باشد که در ۶۰ کلاستر دسته بندی شده اند. در این تحقیق برای هر متن الگوریتم پیشنهادی پیاده سازی شده و سپس خلاصه ی استخراج شده با خلاصه های سیستم های خلاصه ساز ماشینی معرفی شده در DUC2001 مقایسه شده است. همچنین برای هر متن یک خلاصه توسط ابزار SweSum که یک ابزار خلاصه ساز آنلاین تجاری و مؤفق است تهیه کرده ایم. در این ارزیابی نتیجه با خلاصه ی ایجاد شده توسط این ابزار نیز مقایسه شده است. برای ارزیابی از ابزار ROUGE[۱] استفاده نمودیم. در سال های اخیر، در مقالات مختلف از این ابزار برای ارزیابی به دفعات استفاده شده است. سیستم های موجود در DUC هم با این ابزار ارزیابی شده اند.

۲-۳-۱-۳-۴- روش ارزیابی

[۱] - Recall-Oriented Understudy for Gisting Evaluation

به منظور ارزیابی ابتدا معیار F-measure که ترکیبی از دو معیار recall و دقت می باشد، مورد توجه قرار گرفته‌است. میانگین معیار F-measure برای ۱۲ سیستم ماشینی به طور جداگانه و سیستم پیشنهادی، به ازای تمامی 600 متن موجود در DUC2001 در سه معیار ROUGE1، ROUGE2 و ROUGESU4 در شکل شماره ۴-۳ دیده می شود. همانطور که مشاهده می شود در هر سه معیار فوق الذکر عملکرد روش پیشنهادی بهبود یافته است.



شکل ۴-۳- نمودار نتایج ارزیابی الگوریتم پیشنهادی با سیستم SweSum و ۱۱ سیستم موجود در DUC2001 [کام۸۹]

۴-۴-ارزیابی روش وزن دهی پیشنهادی

۴-۴-۱- توصیف داده ها

در این تحقیق، تمامی آزمون های خود را جهت ارزیابی روش وزن دهی پیشنهادی بر روی چهار مجموعه ی سند مشهور در حوزه ی بازیابی اطلاعات انجام داده ایم. این چهار مجموعه Medline، Cranfeild، CISIو CACM می باشند که به معرفی مختصر هر یک خواهیم پرداخت. شایان ذکر است که تمامی آزمون ها جهت بدست آوردن ضرایب مورد استفاده در مراحل محاسبه ی وزن کلمات، تنها بر روی مجموعه ی داده ی Medline صورت گرفته است.

یکی از خصوصیات مشترک چهار مجموعه ی مورد استفاده، نداشتن فرمت XML و یا TREC می باشد. هرچند که برای مجموعه های جدید که معمولاً در دو قالب مذکور تولید می شوند، کدهای استحصال متن نوشته شده است. دومین خصوصیت مشترک همه ی این مجموعه اسناد، وجود تعدادی پرس و جو برای هر کدام است که به صورت انسانی برای آنها تولید شده است. از این جهت، به صورت دقیق، مشخص است که چه اسنادی از یک مجموعه اسناد در قبال یکی از این پرس و جو ها، باید بازگردانده شوند.

کوچکترین مجموعه از این بین، Medline با ۱۰۳۳ سند، ۵۸۳۱ کلمه و ۳۰ پرس و جو استاندارد شده است.

در جدول ۴-۳ اطلاعات آماری درباره ی وضعیت این چهار مجموعه دیده می شود.

جدول ۴-۳- اطلاعات کلی معرفی کننده چهار مجموعه داده ارزیابی

Collection	Number of vectors	Average vector length(number of terms)	Standard deviation of vector length	Average frequency of terms in vectors	Percentage of terms in vectors with frequency 1
MED					
Documents	1033	51.6	22.78	1.54	72.7
Queris	30	10.1	6.03	1.12	90.76
CACM					
Documents	3204	24.52	21.21	1.35	80.93
Queris	64	10.8	6.43	1.15	88.68
CISI					
Documents	1460	46.55	19.38	1.37	80.27
Queris	112	28.29	19.49	1.38	78.36
CRAN					
Documents	1398	53.13	22.53	1.58	69.5
Queris	225	9.17	3.19	1.04	95.69

۲-۴-۴- معیارهای وزن دهی مورد تحقیق

شش معیار با کارایی خوب در حوزه ی مدل های فضای برداری موجود می باشد. البته قطعاً تاکنون تعداد زیادی از معیارهای وزن دهی در این حوزه ابداع شده، که به دلیل پیچیده بودن و بعضاً غیر قابل پیاده سازی بودن آنها، در تحقیقات مورد استناد و ارزیابی قرار نمی گیرند. اما شش معیار خوب، کارا و البته پرکاربرد در این حوزه موجود هستند، که در جدول ۴-۴ به آنها اشاره شده است. در دو حوزه ی مدل های بولی و احتمالی نیز معیارهای خوبی مانند BM25 موجود می باشد. اما طبق گفته ی [SPA00] هیچ کدام از این معیارها برتری نسبت به شش معیار ارائه شده در جدول ۴-۴ ندارند. به این لحاظ و به دلیل وقت گیر بودن پیاده سازی معیاری مانند BM25 از ارزیابی کار با این معیار، صرف نظر شده است.

جدول ۴-۴- شش معیار وزن دهی مشهور به تفکیک پارامترهای وزن محلی و وزن سراسری آنها

Method	Local Weight	Global Weight
TF-IDf	$f_{i,j}$	$\text{Log}\ \frac{N}{n_i}$
LOG-IDF	$1+\log f_{i,j}$ if $f_{i,j}>0$ 0 if $f_{i,j}=0$	$\text{Log}\ \frac{N}{n_i}$
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	$\frac{gf_i}{n_i}$	$\text{Log}\ \frac{N}{n_i}$
No-Weight	$f_{i,j}$	None

در این معیار ها $f_{i,j}$ فرکانس نسبی تکرار کلمه ی i در سند j، N تعداد کل اسناد، n_i تعداد کل اسنادی که کلمه i در آن ها تکرار شده است، gf_i تعداد کل تکرار کلمه ی i در مجموعه ی اسناد و x_j فرکانس ماکزیمم در بین کلمات سند j می باشد.

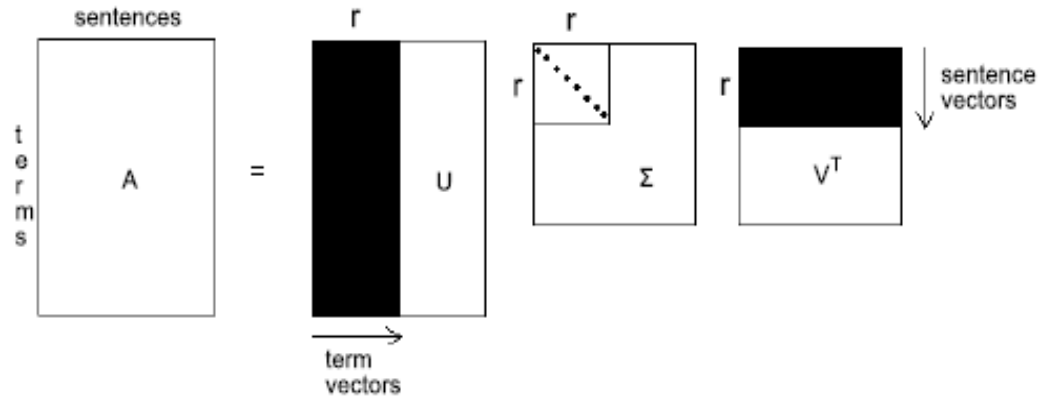
۳-۴-۴- الگوریتم بازیابی مورد استفاده

در این تحقیق از الگوریتم مشهور LSIبرای برگزاری آزمون ها، که تماماً حول تأثیر معیار وزن دهی بر کاربردهای مرتبط با بازیابی اطلاعات می باشند، استفاده شده است. به این منظور در این بخش به توضیح مختصر نحوه ی اعمال این الگوریتم می پردازیم.

LSI در ابتدا برای حل مشکل هم خانواده‌ها و هم آوایی‌ها در بازیابی اطلاعات معرفی شد[Dee 90]. از آن پس LSI در کانون توجه محققان در زمینه پردازش متن قرار گرفت و در بسیاری از مقالات به صورت تئوری و عملی آنالیز گردید[Dee 90][Pap 00]. LSA در درون خود از روش SVD[1] استفاده می‌کند. SVD ماتریس A با ابعاد

[1] - singular value decomposition

m×n را به سه ماتریس $A = USV^T$ تجزیه می‌کند که $U = [u_{ij}]$ ماتریس ارتونرمال[1]ستونی m×m بوده و ستون‌های آن بردارهای یکه سمت چپ نامیده می شوند؛ $S = diagonal(\sigma_1, \sigma_2, \dots, \sigma_n)$ ماتریس قطری n×n است که عناصرقطر اصلی آن مقادیر ویژه غیرمنفی می‌باشند که به صورت نزولی مرتبط شده‌اند و $V = [v_{ij}]$ هم ماتریس ارتونرمال سطری بوده و ستون‌های آن بردارهای یکه سمت راست نامیده می‌شوند. خروجی الگوریتم SVD در شکل ۴-۴ قابل مشاهده است.



شکل ۴-۴- خروجی SVD برای ماتریس کلمه-جمله

اگر rank(A)= r باشد آنگاه حالت زیر برقرار خواهد بود :

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq \sigma_{r+1} = \dots = \sigma_n = 0$$

از دیدگاه ریاضی، استفاده از SVD برروی یک ماتریس منجر به نگاشت فضای m بعدی به یک فضای r بعدی می شود و در بسیاری از کاربردها از آن برای کاهش ابعاد در ماتریس‌های تنک استفاده می شود. از دیدگاه پردازش زبان طبیعی، SVD باعث استخراج روابط معنایی پنهان در ساختار داده می‌شود.

۴-۴-۴- آزمون ها

برای بررسی دقیق نحوه و میزان تأثیر گذاری روش پیشنهادی، آزمون های متنوعی حول کاربرد بازیابی اطلاعات طراحی شده است. در این تحقیق، دو آزمون متداول تر مورد بررسی قرار گرفته است. همچنین، یک آزمون جدید برای بررسی میزان قدرتمندی یک روش وزن دهی به کلمات ابداع و معرفی شده است.

۱-۴-۴-۴- آزمون انتخاب ابعاد حقیقی

[1]- orthonormal

در این آزمون به تأثیر روش های وزن دهی در فرآیندی به نام انتخاب ابعاد حقیقی[1]می پردازیم. همانطور که بیان شد ماتریس جواب الگوریتم LSI به تعداد کلمات بعد دارد. اما همه ی این ابعاد باارزش نیستند و یا به عبارتی جزء مفاهیم مجموعه ی داده ی مورد نظر نمی باشند. برای انتخاب تعدادی از این ابعاد که باارزش می باشند، روش هایی موسوم به انتخاب ابعاد حقیقی ارائه شده است.

یک حدس شهودی که می تواند کارآیی الگوریتم بازیابی مانند LSI را تحت تأثیر قرار دهد، تعداد ابعاد حقیقی است که در حین انجام عمل کاهش ابعاد، برگردانده می شوند. این انتخاب ها باید به قدر کافی بزرگ باشند تا بتوانند به خوبی خصوصیات مجموعه ی داده را توصیف کنند. همینطور این انتخاب ها باید به قدر کافی، به منظور عدم وارد شدن اطلاعات بی ارزش در مجموعه ی داده، کوچک باشند. همانطور که دیدیم در الگوریتم LSI تنها ماتریس اول U به عنوان ماتریس جواب مورد پردازش قرار می گیرد.

محققین تاکنون تعداد کمی از روش ها را برای انتخاب ابعاد را ارائه کرده اند. در [ZHU06] روشی برای ایجاد یک مدل واضح برای همه مقادیر ماتریس جواب ها و تخمین جایگاه گپ با حداکثر سازی یک تابع شباهت همسایگی[2](PL) ارائه شده است. اندیسی از ماتریس جواب که ماکزیمم مقدار PL در آن رخ داده است، به عنوان تعداد ابعاد حقیقی، انتخاب می شود. تحلیل موازی[3](PA) [EFR05] یک روال نمونه گیری دوباره است که در آن مقادیر ویژه از داده های تحقیق با مقادیر ویژه یک ماتریس اتفاقی که ابعادی برابر با داده های اصلی دارد، مقایسه می شود. تعداد بردارهای ویژه ای که از ماتریس اول برداشته می شود، برابر است با تعداد مقادیر ویژه ای از ماتریس اول که بزرگتر از متوسط مقادیر ویژه ی ماتریس اتفاقی است. طول توصیف مینیمم اصلاح شده[4](AMDL) یک روش برای اندازه گیری میزان نزدیکی مقادیر ویژه به وسیله چک کردن نسبت میانگین حسابی و درجه ی دوم می باشد که به تشخیص ابعاد حقیقی منجر می شود [KUM08]. اندیسی از ماتریس جواب که مینیمم مقدار تابع AMDL را داراست، به عنوان تعداد ابعاد حقیقیی برگردانده می شود.

۱-۱-۴-۴-۴- نتایج تجربی

[1] - Intrinsic dimensionality selection
[2] - Profile Likelihood
[3] - Parallel Analysis
[4] - Amended Minimum Description Length

برای نشان دادن تأثیر معیار های وزن دهی بر روش های انتخاب ابعاد حقیقی، آزمایش هایی را بر روی مجموعه داده های Medline، Cranfeild، CISI و CACM طراحی کرده ایم. در این آزمون تأثیر معیار وزن دهی بر هر سه روش PL، PA و AMDL را بررسی می کنیم.

بر اساس استراتژی که معیار وزن دهی بر آن استوار است، میزان اهمیت یک کلمه افزایش و یا کاهش می یابد [ERI99] و [DEB04]. به عنوان مثال معیار TF-IDF بیشترین وزن را به کلماتی می دهد که به طور متناوب در تعداد کمی از اسناد از مجموعه ی همه ی اسناد تکرار شده اند. برای مقایسه و نشان دادن تأثیر معیار وزن دهی بر اهمیت کلمات، در جدول ۴-۵ مقادیر نسبت داده شده به کلمه ی "Bacillus" در مجموعه داده ی Med، بوسیله ی معیار های وزن دهی جدول ۴-۴ نشان داده شده اند. همانطور که مشاهده می شود، مقادیر تخصیص داده شده توسط روش پیشنهادی به کلمه مورد نظر در اسناد مختلف از پرا کندگی پایینتری نسبت به مقادیر اختصاص داده شده به همان کلمه توسط معیارهای وطن دهی شش گانه برخوردار است. این یکی از دلایل واضح توانایی معیار پیشنهادی در خوشه بندی اولیه ی داده ها می باشد. حین عمل بازیابی برای کلمه مورد نظر سند های موجود در جدول ۴-۵ باید بازگردانده شوند، که به دلیل نزدیک تر بودن وزن اختصاص داده شده توسط معیار پیشنهادی، احتمالاً بازگشت تعداد زیادی از این اسناد بسیار بالا می باشد.

جدول ۴-۵- وزن اختصاص داده شده به کلمه "Bacillus " توسط معیارهای مختلف در اسنادی که طی عمل بازیابی اطلاعات برای این کلمه بازگردانده می شوند

Document No	No-weight	TF-IDF	LOG-IDF	TF-ENPY	ATF1-IDFP	GF-IDF	Proposed Approach
22	1	3.8957	3.8957	4.4367	2.9854	0.4730	5.8753
144	3	11.6871	8.1756	7.5478	8.5986	0.9461	7.2464
145	1	3.8957	3.8957	4.4367	2.9854	0.4730	4.4532
146	2	3.8957	3.8957	4.4367	2.9854	0.4730	6.0075
147	2	7.7914	6.5960	5.8790	5.3985	0.7498	6.2504
194	4	7.7914	6.5960	5.8790	5.3985	0.7498	7.0092
195	3	15.5828	9.2963	9.5683	8.9871	1.0984	9.7903
196	1	11.6871	8.1756	9.2904	7.4592	0.9461	7.1398
197	2	3.8957	3.8957	4.4367	2.9854	0.4730	5.3067
198	2	7.7914	6.5960	5.8790	5.3985	0.7498	5.2740
199	1	7.7914	6.5960	5.8790	5.3985	0.7498	6.1003
473	1	3.8957	3.8957	4.4367	2.9854	0.4730	4.9845
483	1	3.8957	3.8957	4.4367	2.9854	0.4730	3.7691

همانطور که می بینیم، در جدول بالا یک شهود فوری از بهبود کارایی روش پیشنهادی، مشاهده می گردد. از آنجاییکه در الگوریتمی مانند LSI در رابطه با یک کلمه، اسنادی که بالاترین وزن آن کلمه را در مجموعه ی اسناد در خود دراند، انتخاب می شوند، در مورد کلمه ی ''Bacillus'' نزدیکی وزن میان اسناد ارائه شده در این جدول برای معیار پیشنهادی ما، باعث انتخاب و برگرداندن بسیاری از این اسناد می گردد.

شکل های ۴-۵، ۴-۶ مقادیر تولید شده برای توابع انتخاب ابعاد حقیقی PL و AMDL بر روی مجموعه ی داده MED و به ازای هر معیار وزن دهی را نشان می دهند. اشکال مرتبط با نتایج حاصل از توابع PL و AMDL بر روی سه مجموعه سند دیگر و تابع PA بر روی هر چهار مجموعه سند به دلیل بالا بودن تعداد آنها، در ضمیمه ج ارائه شده است. اعمال هر معیار وزن دهی برای ساخت ماتریس کلمه-سند منجر به تغییرات گوناگون و در نتیجه تولید تعداد ابعاد حقیقی متفاوت می شود. تعداد ابعاد حقیقی انتخاب شده ی تقریبی برای هر معیار وزن دهی به همراه روش انتخاب ابعاد حقیقی در جداول ۴-۶، ۴-۷ و ۴-۸ دیده می شود.

جدول ۴-۶- مقادیر ابعاد حقیقی برای چهار مجموعه داده توسط تابع PL به تفکیک معیار وزن دهی

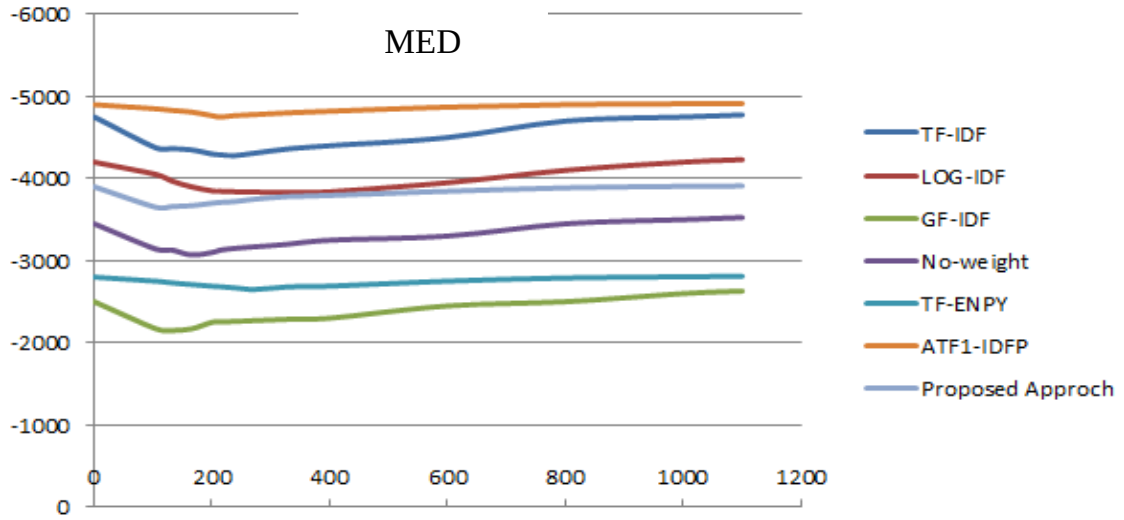
Method	MED	CRAN	CACM	CISI
No-weight	165	158	430	195
TF-IDF	238	296	739	398
LOG-IDF	327	406	927	501
TF-ENPY	273	252	854	344
ATF1-IDFP	215	290	903	387
GF-IDF	132	105	415	151
Proposed Approach	105	96	293	114

جدول ۴-۷- مقادیر ابعاد حقیقی برای چهار مجموعه داده توسط تابع AMDL به تفکیک معیار وزن دهی

Method	MED	CRAN	CACM	CISI
No-weight	103	105	88	96
TF-IDF	69	54	48	40
LOG-IDF	27	28	27	19
TF-ENPY	73	74	65	52
ATF1-IDFP	68	53	74	45
GF-IDF	103	118	93	98
Proposed Approach	19	21	23	19

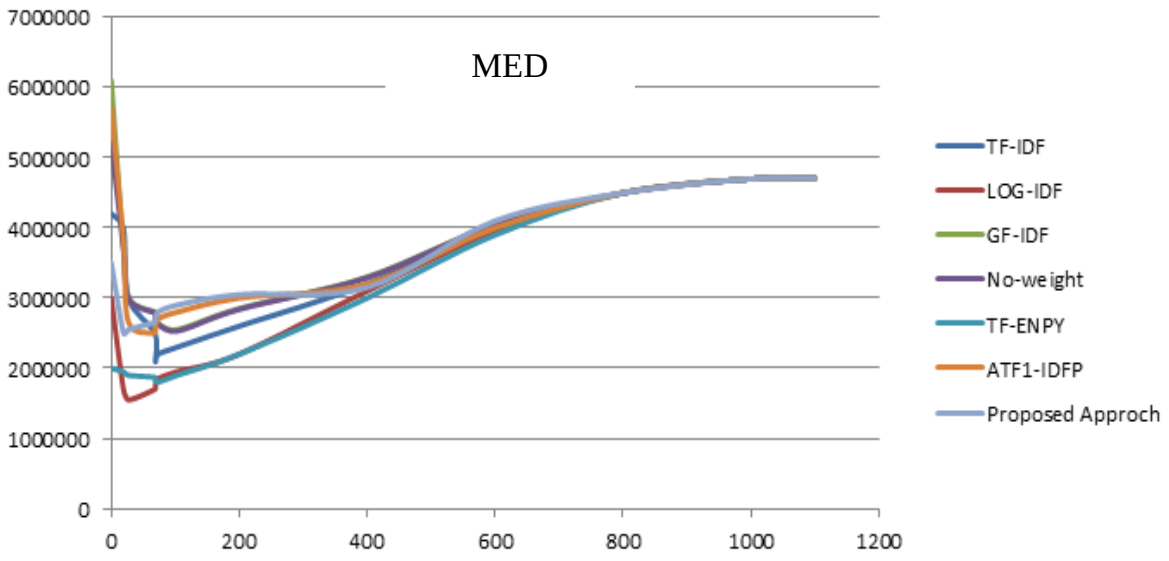
جدول ۴-۸- مقادیر ابعاد حقیقی برای چهار مجموعه داده توسط تابع PA به تفکیک معیار وزن دهی

Method	MED	CRAN	CACM	CISI
No-weight	106	118	134	135
TF-IDF	124	164	244	223
LOG-IDF	424	197	292	253
TF-ENPY	108	115	201	187
ATF1-IDFP	173	214	223	134
GF-IDF	102	105	181	125
Proposed Approach	103	98	134	106



محور x ها: تعداد ابعاد (Number of Dimensions) و محور y ها: مقادیر شباهت همسایگی (Likelihood Value)

شکل ۴-۵- نمودار مقادیر ابعاد حقیقی بازگردانده شده برای مجموعه داده MED به ازای معیارهای وزن دهی مختلف توسط تابع PL



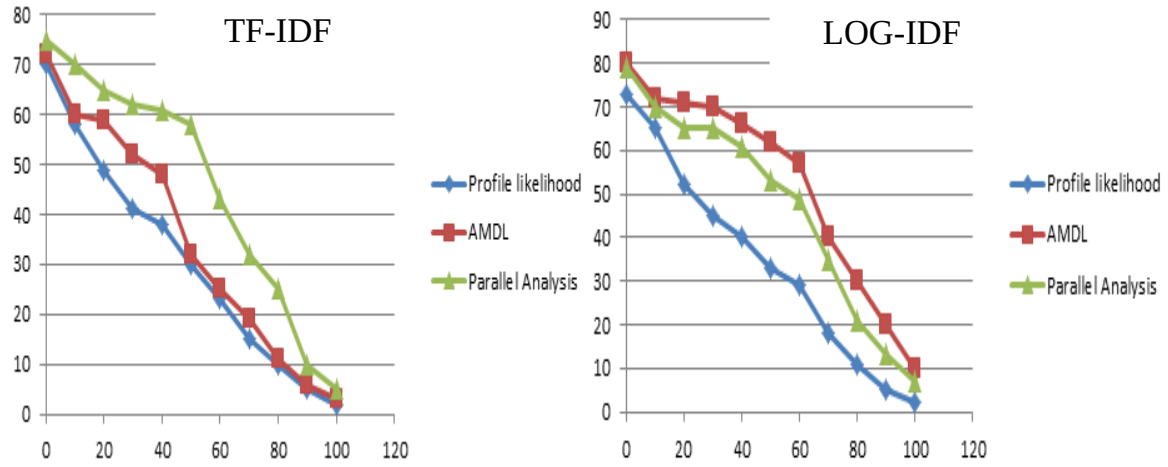
محور x ها: تعداد ابعاد (Number of Dimensions) و محور y ها: مقادیر تابع AMDL

شکل ۴-۶- نمودارهای مقادیر ابعاد حقیقی بازگردانده شده برای مجموعه داده MED به ازای معیارهای وزن دهی مختلف توسط تابع AMDL

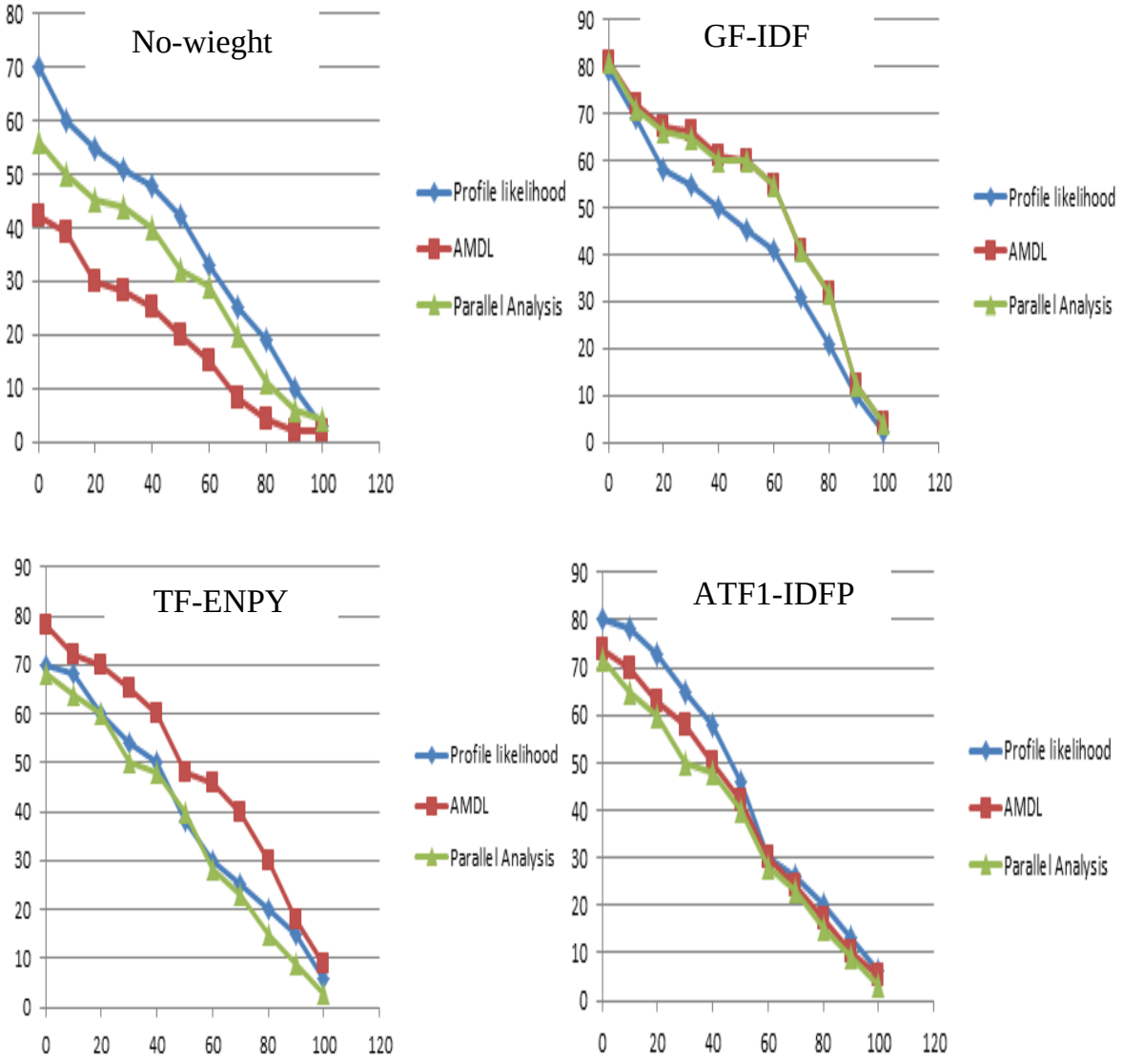
نتایج بدست آمده، ثابت می کند که معیارهای وزن دهی تأثیر بسزایی در تعداد ابعاد حقیقی دارند. همانطور که مشاهده می شود، برای تمامی روشهای انتخاب ابعاد حقیقی و بر روی همه ی مجموعه های اسناد، مقدار ابعاد حقیقی بازگردانده شده برای معیار وزن دهی پیشنهادی، کمتر است. کم بودن این مقدار به این معنا است که تعداد کلاسترهای ایجاد شده و یا در حقیقیت تعداد مفاهیم استخراج شده برای یک مجموعه ی سند کمتر می باشد. از آنجایی که هر یک از مجموعه ی اسناد، در موضوع خاصی قرار دارند، این بدیهی است که تعداد مفاهیم موجود در مجموعه ی اسناد، نباید عدد بزرگی را نشان دهد. در عین حال بزرگ بودن تعداد ابعاد حقیقی بازگردانده شده، دو معنای مختلف را داراست: ۱- عملکرد بد روش انتخاب ابعاد حقیقی در تشخیص درست این ابعاد و یا ۲- توزیع نامناسب وزن دهی به کلمات توسط روش وزن دهی و در نتیجه افزایش تعداد مفاهیم به صورت غلط.

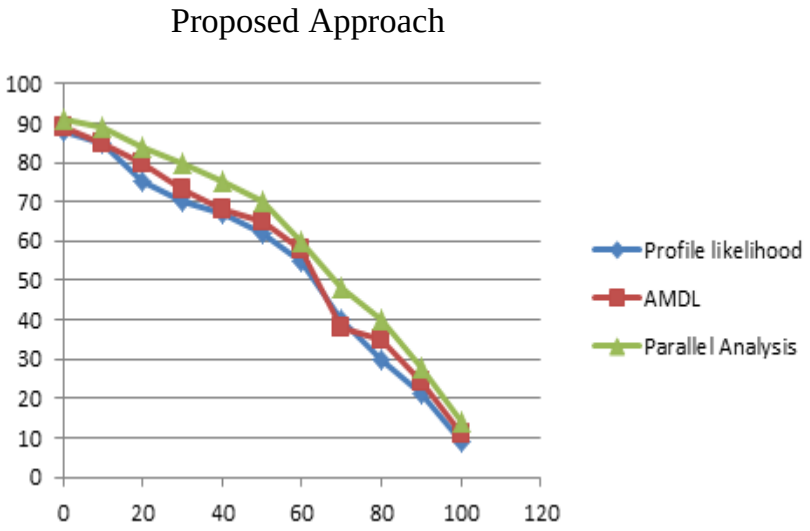
همانطور که بعداً خواهیم دید، در صورت بروز مشکل دوم، بازیابی اطلاعات به لحاظ کیفی افت شدیدی را شاهد خواهد بود. در مورد روش پیشنهادی، کم بودن تعداد ابعاد حقیقی انتخاب شده، در صورت بالا بودن میزان دقت (که در آزمون بعدی بالا بودن آن تأیید می شود)، نشان دهنده ی تأثیر آن در خوشه بندی اولیه ی داده ها و تشخیص ضمنی مفاهیم موجود در مجموعه های اسناد می باشد.

از آنجایی که تعداد کلمات یکتا و تعداد کلمات در هر سند، برای هر مجموعه سند متفاوت است، کارایی هر یک از معیارهای وزن دهی نیز متفاوت می شود. در ادامه، در مورد اینکه چگونه تأثیر معیار وزن دهی در انتخاب ابعاد حقیقی منجر به تأثیر در کیفیت بازیابی اطلاعات می شود، بحث خواهیم کرد. برای اثبات این مسئله، اقدام به طراحی آزمایش هایی بر روی بازیابی اطلاعات بر روی مجموعه های اسناد و اندازه گیری مقدار دقت الحاقی[1] در سطوح استانداردی از Recall خواهیم کرد [KUM09]. دقت الحاقی به معنای ماکزیمم دقت به ازای همه ی Recall های r' که بزرگتر از Recall در سطح r هستند می باشد. در اینجا برای کل ۳۰ پرس و جو موجود بر روی مجموعه ی سند MED تنها یک دقت و Recall را ارائه می کنیم. به این صورت که از میان کل اسناد بازگردانده شده برای کل پرس و جو ها، مقادیر دقت و Recall مشخص می شوند. شکل ۴-۷ منحنی های دقت الحاقی بر روی مجموعه سند Medline را نشان می دهد. برای هر معیار وزن دهی، گراف ها نشان دهنده ی منحنی های دقت الحاقی بدست آمده توسط اعمال مدل LSI به همراه ابعاد حقیقی بدست آمده توسط روش های مختلف تقریب انتخاب این ابعاد می باشند.



[1] - Interpolated precision


No-wieght
GF-IDF
TF-ENPY
ATF1-IDFP
Profile likelihood
AMDL
Parallel Analysis



محور x ها: سطوح استاندارد Recall و محور y ها: مقادیر دقت الحاقی (Interpolatedدقت)

شکل ۴-۷- نمودارهای افزایش کارایی الگوریتم LSI بر اساس معیارهای مختلف وزن دهی و توابع انتخاب ابعاد حقیقی متفاوت

همانطور که می بینیم، زمانی که معیار وزن دهی پیشنهادی مورد استفاده قرار می گیرد، LSI به همراه روش انتخاب ابعاد حقیقی PA بهترین کارایی بازیابی را در همه ی سطوح Recall نشان می دهد. همچنین LSI به همراه روش انتخاب ابعاد حقیقی PL بدترین دقت الحاقی را در همه ی سطوح Recall از خود نشان می دهد. زمانی که از معیار TF-IDF استفاده می شود، LSI به همراه روش انتخاب ابعاد حقیقی PA بهترین دقت الحاقی در همه ی سطوح Recall از خود نشان می دهد. در زمان استفاده از معیار GF-IDF، دو روش PA و AMDL به ترتیب ۱۰۲ و ۱۰۳ بعد را به عنوان ابعاد حقیقی برمی گردانند. در عین حال می بینیم که کارایی LSI برای این دو روش و این معیار وزن دهی برابر می باشد. زمانی که از معیار LOG-IDF استفاده می شود، LSI به همراه روش انتخاب ابعاد حقیقی PL بهترین دقت الحاقی در همه ی سطوح Recall و با استفاده از روش انتخاب ابعاد حقیقی AMDL کارایی ضعیفی را از خود نشان می دهد. همینطور می بینیم که علیرغم اینکه هنگام استفاده از روش پیشنهادی ابعاد حقیقی انتخاب شده توسط روش های PA و PL به ترتیب ۱۰۵ و ۱۰۳ می باشد، اما با استفاده از LSI میزان دقت الحاقی در همه ی سطوح Recall برای این دو روش برابر نمی باشد. هر چند که فاصله ی میان دو نمودار برای آنها، کم می باشد.

یکی دیگر از نکات قابل مشاهده از نمودارهای شکل ۴-۷ نزدیک بودن نمودارهای سه گانه توابع انتخاب ابعاد حقیقی برای روش پیشنهادی است. این نشان دهنده پایداری بالاتر وزن دهای داده شده در افزایش دقت است. به

این ترتیب، دقت حاصله بر روی بازیابی اطلاعات در مجموعه داده های مختلف، وابستگی چندانی به تابع انتخاب ابعاد حقیقی ندارد. بار اصلی این وابستگی توسط یک وزن دهی خوب، به سمت خوشه بندی و پیش خوشه بندی منتقل شده است. این نکته نشان دهنده قدرت بالاتر روش پیشنهادی در پیش خوشه بندی خوب مجموعه داده ها است. همانظور که از شکل ۴-۷ مشاهده می شود، برای معیارهای وزن دهی دیگر فاصله میان نمودارهای سه گانه از نظم و ترتیب خاصی برخوردار نیست.

دو نکته ی واضح در مورد شکل ۴-۷ به این ترتیب است که: ۱- روش پیشنهادی باعث افزایش میزان دقت در تمامی سطوح recall و برای هر سه نوع روش انتخاب ابعاد حقیقی شده است. این نکته به تنهایی نشان دهنده ی میزان تأثیر روش پیشنهادی در افزایش کارایی روشهای انتخاب ابعاد حقیق است. و ۲- اینکه فاصله ی میان نمودارهای سه روش انتخاب ابعاد حقیقی برای روش پیشنهادی بسیار کم است. این نکته نشان دهنده ی تأثیر بالای روش پیشنهادی در خوشه بندی مقدماتی کلمات و هر چه بی نیاز تر کردن نتیجه ی LSI به استفاده از مدل خاصی برای انتخاب ابعاد حقیقی می باشد. همانطور که مشاهده می شود، هیچ کدام از روش های وزن دهی دیگر، چنین قابلیتی را از خود بروز نداده اند.

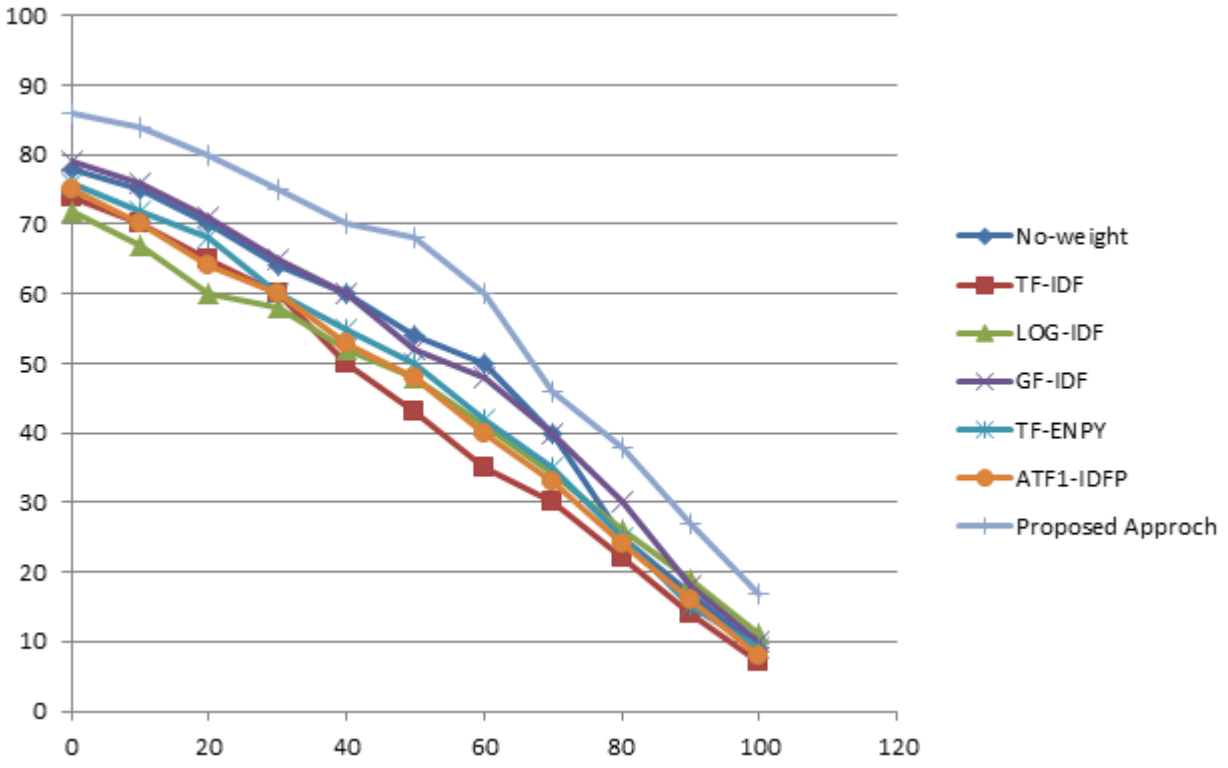
۲-۴-۴-۴- آزمون MAP

MAP[1] یکی از آزمون های معمول در بازیابی اطلاعات اعمال الگوریتمی مانند LSI بر روی مجموعه ی اسناد و محاسبه ی دقت، Recall و F-measure می باشد. در این آزمون، از هیچ روش انتخاب ابعاد حقیقی استفاده نمی کنیم و تنها به گام های معمول الگوریتم LSI اکتفا می کنیم.

دو معیار ابتدایی که در این آزمون به ارائه ی آن خواهیم پرداخت، معیار Recall و دقت می باشند. نکته ی قابل توجه در روش های بازیابی اطلاعات، بالا بودن میزان معیار Recall آنها می باشد. به نحوی که روشی مانند PLSI دارای Recall برابر ۹۵٪ در تمامی وضعیت ها می باشد. اما همانطور که در شکل ۴-۸ مشاهده می کنیم، در همه ی معیارهای وزن دهی، میزان دقت در نقاطی که Recall بالایی وجود دارد، بسیار پایین است. در حالیکه نقطه ی قوت معیار وزن دهی پیشنهادی ما، در افزایش مقدار دقت با حفظ Recall می باشد، که در شکل ۴-۸ مشاهده می شود.

[1] - Mean Average Precision

برای نشان دادن معیار دقت برای تعداد زیادی از پرس و جو ها بر روی هر یک از مجموعه ی داده ها مبتنی بر گستره ای از Recall ها، به ارائه ی یک نمونه ی دیگر از این معیارهای ارزیابی که میانگین میانگین دقت (MAP) می باشد، می پردازیم. برای هر پرس و جو یک بازه از دقت را داریم. در صورتی که از این بازه میانگین گرفته شده و سپس میان پرس و جو های مختلف بر روی یک مجموعه سند نیز، میانگین گرفته شود، آنرا MAP نامند. حال می توانیم این معیار را به یک سطح بالاتر نیز ارتقاء دهیم و از نتایج حاصل بر روی هر یک از مجموعه ی داده ها نیز یکبار دیگر میانگین بگیریم.



محور x ها: سطوح مختلف Recall و محور y ها: میانگین میانگین دقت به دست آمده برای هر چهار مجموعه داده

شکل ۴-۸- نمودار معیار MAP برای چهار مجموعه داده با هم به ازای معیار های متفاوت وزن دهی

همانطور که در شکل ۴-۸ مشاهده می شود، روش پیشنهادی از بالاترین میزان دقت در همه ی سطوح recall برخوردار است. یکی از نقاط قوت معیار MAP، توانایی آن در نشان دادن عدم وابستگی یک معیار وزن دهی به مجموعه ی سند خاصی می باشد. زیرا این معیار بر روی میانگینی از نتایج مربوط به مجموعه های مختلف بنا شده است.

۳-۴-۴-۴- آزمون انحراف از معیار کلمات وزن داده شده

همانطور که دیدیم، در هر دو آزمون گذشته، کارایی روش پیشنهادی در بهبود خوشه بندی اسناد، مشخص شد. در ادامه ی سلسله تحقیقات، به عنوان آزمونی که مشخص کننده ی یک ویژگی ذاتی برای روش های وزن دهی باشد، توجه ما به معیار انحراف از معیار معطوف شد. بر اساس آزمونی که پیش رو داریم و با توجه به نتایج به دست آمده در دو آزمون گذشته و اثبات تجربی این موضوع که روش پیشنهادی کارایی بهتری نسبت به روش های موجود دارد، می توان نتیجه گرفت که انحراف از معیار نیز، معیار خوبی برای ارزیابی میان روش های وزن دهی به کلمات می باشد. این آزمون نیز بر روی هر چهار مجموعه ی سند صورت گرفته است.

در این آزمون از تابع توزیع نرمال که در رابطه ی ..... مشاهده می شود استفاده شده است. متغیرهای تصادفی در این تابع، وزن های اختصاص داده شده به کلمات که در فضای اعداد حقیقی بزرگتر از صفر هستند می باشد. همچنین فضای رخداد متغیرهای تصادفی خود کلمات می باشد. در این آزمون، ابتدا به دسته بندی کلمات بر اساس مقدار وزن داده شده به آنها می پردازیم. به این صورت که اگر تعداد کلمات ۱۰۰ و تعداد کلمات با وزن ۲، چهار کلمه باشند، آنگاه $p(2) = 0.04$. سپس با استفاده از احتمالات وزن های موجود، بر اساس تابع توزیع نرمال رابطه ..... به ترسیم نمودار این تابع می پردازیم.

همانطور که در اشکال ..... مشاهده می شود، روش پیشنهادی از انحراف معیار بزرگتری، نسبت به سایر معیارهای وزن دهی به کلمات برخوردار است. یک توجیه مبتنی بر نتیجه از این نمودارها این است که، در انحراف معیار بزرگتر، توزیع وزن کلمات در گستره ی پهن تری صورت گرفته و از اجتماع کلمات در ناحیه ی مشخصی جلوگیری می شود. اجتماع بیش از حد کلمات در یک ناحیه ی مشخص، باعث بازگرداندن اشتباه اسناد در ارتباط با پرس و جو ها و بالا رفتن میزان خطا در ساخت مفاهیم می شود. توزیع کلمات در بازه ی پهن تر امکان مقایسه پذیری کلمات با یکدیگر را بالاتر می برد. همینطور این نوع توزیع باعث کاهش مقایسه پذیری دسته های به وجود آمده از کلمات با یکدیگر و به نوعی پایداری در خوشه بندی می شود.

۵-۴- خلاصه

در این فصل ابتدا به معرفی تعدادی از ابزاری که برای پیاده سازی روش پیشنهادی مورد استفاده قرار گرفته اند، پرداختیم. در بعضی از گام ها نیز به پیاده سازی های جزئی شخصی اکتفا شده است. البته اکثر این پیاده سازی ها تنها برای قسمت پیش پردازش های زبانی مورد استفاده قرار می گیرند.

سپس به ارائه نتایج مرتبط با اثبات تجربی حدس های شهودی درباره ی انتقال های معنایی نظریه مرکزیت پرداخته ایم. این نتیجه یک کار تجربی بر مبنای نظریه مرکزیت حول خلاصه سازی تک سنده متون می باشد.

سپس به ارائه نتایج مرتبط با ارزیابی روش پیشنهادی با شش معیار وزن دهی مشهور پرداخته ایم. ابتدا یک معرفی اجمالی از چهار مجموعه داده استاندارد برای کاربرد بازیابی اطلاعات کرده ایم. سپس به معرفی شش معیار وزن دهی مشهور و الگوریتم مشهور بازیابی اطلاعات LSI پرداخته ایم.

در انتها به معرفی سه آزمون برای مشاهده تأثیرگذاری روش پیشنهادی بر کاربرد بازیابی اطلاعات پرداخته ایم. به همراه این معرفی برای هر آزمون نمودارهای نتایج مرتبط با آن نیز به نمایش گذاشته شده است.

نتایج حاصله همگی از افزایش میزان دقت کاربرد بازیابی اطلاعات در هر چهار مجموعه داده حکایت دارند.

فصل پنجم:

نتیجه گیری و کارهای آتی

۵– نتیجه گیری و کارهای آتی

۵–۱– نتیجه گیری

در این پایان نامه یک روش جدید برای وزن دهی به کلمات در کاربردهای پردازش زبان، بازیابی اطلاعات و ... ارائه شد. این روش جدید بر مبنای دو دیدگاه معنایی و زبانی بنا شده است. یکی از مهمترین دلایل توجه به چنین دیدگاه هایی در وزن دهی به کلمات، وابستگی ویژگی های تعریف کننده متن، یعنی کلمات، به خواصی ذاتی و ضمنی مانند معنای کلمه و نقش گرامری آن می باشد. همانطور که دیدیم، بسیاری از روش های کنونی وزن دهی به کلمات، تنها بر خواص آماری کلمات در متون، که همان فرکانس تکرار آنها می باشند، تکیه دارند. این وابستگی به تنهایی باعث عدم کارایی چندان مناسبی در خوشه بندی اولیه کلمات می شود، که به وضوح در نتایج بدست آمده میان روش های مختلف، مشاهده می شود.

برای پرکردن این خلأ – توجه به ویژگی های ذاتی زبان- در این پایان نامه با انتخاب روش وزن دهی TF-IDF و تغییر پارامتر وزن دهی محلی TF از دیدگاه معنایی و زبانی، سعی شده است که، تا حد زیادی کارایی این روش بهبود یابد. دلیل اصلی انتخاب روشی مانند TF-IDF به عنوان معیار مبنا، سادگی پیاده سازی و محاسبه ی آن می باشد. همین دو دلیل، علت اصلی محبوبیت این روش، در میان معیارهای وزن دهی نیز می باشد. در این مسیر دو ایده ی اصلی برای پوشش دادن همزمان معنا و گرامر زبان مورد بررسی قرار گرفت.

برای پوشش دادن معنا و اضافه کردن ویژگی های معنایی متن در وزن دهی به کلمات، یک تئوری زبانشناسی به نام نظریه ی مرکزیت مورد تحقیق و استفاده قرار گرفت. یکی از خاصیت های اصلی و اثبات شده ی نظریه ی مرکزیت تاکنون، توانایی آن در تشخیص انسجام و برجستگی محلی در یک متن می باشد. ادعای اصلی این نظریه در تشخیص انتقال های معنایی میان جملات یک متن می باشد. از این رو در پایان نامه ی ذکر شده، مبنای کار بر تغییر پارامتر TF که یک پارامتر وزن محلی می باشد، گذاشته شد. سپس بر اساس چهار شهود ارائه شده، یک اولویت میان چهار انتقال معنایی تعریف شده در نظریه ی مرکزیت قائل شدیم. این اولویت در رابطه ۵-۱ مشاهده می شود.

$$\text{Continue} > \text{Smooth-shift} > \text{Retain} > \text{Rough-shift} \qquad (۵-۱)$$

بر این اساس دو زیر گام پیموده شد. در زیرگام اول بر اساس میانگین تکرار هر انتقال معنایی در مجموعه ای از متون مورد پردازش، وزنی به هر انتقال معنایی داده شده و سپس این وزن در وزن کلمات جمله ای که انتقال معنایی خاصی در آن رخ داده است ضرب شد. نتایج حاصله از این زیرگام چندان مطلوب نبود و به همین دلیل

نحوه ی محاسبه ی وزن هر انتقال معنایی از طریق زیر گام بعدی مورد توجه قرار گرفت. در زیر گام دوم وزن هر انتقال معنایی در یک متن بر اساس فرمول ارائه شده محاسبه شد. در این وزن، ضریبی که برای هر مجموعه ی متن یکبار محاسبه می شود نیز، ضرب شده است. سپس برای هر کلمه، عدد یک که به معنای یکبار دیده شدن آن در جایی از متن است، در وزن انتقال معنایی جمله ی دربرگیرنده ی آن کلمه، ضرب می شود.

برای در بر گرفتن خاصیت های زبانی که عمده ترین آنها گرامر می باشد، با یک حدس شهودی، نقش گرامری هر کلمه در یک متن مورد توجه قرار گرفت. بر این اساس یک اولویت میان نقش های گرامری از جهت اهمیت محلی آنها برقرار شد، که بر این اساس به هر نقش گرامری ضریبی اطلاق می گردد. این اولویت در رابطه ۵-۲ مشاهده می شود.

فاعل > مفعول غیر مستقیم > مفعول مستقیم > قید مکان > .... (۵-۲)

بر این اساس برای هر کلمه وزن آن در هر نقش خاص در یک متن محاسبه شده و میان وزن های بدست آمده در نقش های مختلف میانگین گرفته شد. از این پس وزن نهایی بدست آمده در وزن محاسبه شده ی یک کلمه تا مرحله ی قبل به ازای دیده شدن آن در هرجای متن و در هر نقش گرامری، ضرب می شود.

در ادامه ی توجه به ویژگی ذاتی کلمات در زبان که نقش گرامری آنها است، یک تجربه و شهود در مورد وزن دهی به یک کلمه بر اساس توزیع آن بر حسب نقش های گرامری خاص آن در مجموعه ی متون مورد بررسی قرار گرفت. یکی از نقاط کمبود روش پیشنهادی در این پایان نامه تا این گام عدم توجه به توزیع سراسری کلمه در مجموعه ی متون می باشد. هرچند که در این گام نیز به تغییر پارامتر وزن محلی TF می پردازیم، اما همین پرداختن به توزیع سراسری یک کلمه در نقش های گرامری خاص در سراسر مجموعه ی اسناد، تا حدی چالش مطرح شده را حل می کند. در این گام، به مانند گام قبلی، به محاسبه ی وزن هر کلمه در یک نقش گرامری خاص می پردازیم. سپس برای یک کلمه در نقش های گرامری متفاوتش میانگین گرفته شده و در وزن بدست آمده برای هر کلمه تا این گام ضرب می شود. البته در این گام به دلیل عدم توجیه و شهود تئوریک مناسب هیچ اولویتی میان نقش های گرامری مختلف برقرار نشده است.

در نهایت برای محاسبه ی وزن TF-IDF به ازای هر بار دیده شدن یک کلمه وزن بدست آمده ی نهایی برای آن را جمع می کنیم. در معیار TF به ازای هر بار دیده شدن یکبار یک کلمه یک واحد به مقدار TF اضافه می شود و در نهایت بر تعداد کل کلمات متن تقسیم می شود. در اینجا نیز پس از جمع کل وزن های دیده شده بر کلمات، این

مقدار بر تعداد کل کلمات متن تقسیم می شود. در صورتی که این مقدار بر مجموع وزن کلمات متن تقسیم می شد، تأثیر مقدارهای محاسبه شده کاهش شدیدی پیدا می کرد.

نتایج بدست آمده بر اساس سه آزمون انتخاب ابعاد حقیقی، آزمون MAP و آزمون انحراف از معیار نشان دهنده ی افزایش محسوس کارایی روش پیشنهادی نسبت به شش معیار وزن دهی مشهور بر روی چهار مجموعه ی سند می باشد.

همانطور که دیدیم روش پیشنهادی دو دسته از چالش ها را تا حد زیادی حل کرده است:

۱. چالشهای پیش روی روش های آماری که عدم توجه به معنا و ویژگی های زبانی و در نتیجه پایین بودن میزان دقت و کارایی آنها می باشد.
۲. چالشهای سه گانه پیش رو روش هایی که دید معناگرا به امر وزن دهی به کلمات دارند. از جمله اینکه روش پیشنهادی به هیچ مجموعه دانش اولیه ای وابسته نیست، پردازش های اضافی به دلیل استفاده از روش های یادگیری و یا استفاده از لغتنامه ها و .... ندارد و به مجموعه ی داده ی اولیه وفادار بوده و آنرا تغییر، بسط و گسترش نمی دهد.

چالش بزرگ در مورد روش پیشنهادی زمان اجرای آن می باشد. هرچند که روش های بازیابی مشهور مانند LSI، PLSI و LDA از مرتبه های زمانی بالایی برخوردارند، اما اضافه شدن زمان اجرای جدید برای پردازش های مرتبط با تشخیص هم مرجعی ها، برچسب زنی نقش گرامری و تشخیص انتقال های معنایی باعث کاهش محبوبیت روش پیشنهادی به لحاظ مرتبه زمانی می شود. کمترین میزان زمان اجرا از آن روش LSI با مرتبه ی زمانی O(mlog n) می باشد که m تعداد کلمات و n تعداد اسناد مجموعه ی متون می باشد. هر چند که زمان اجرای اضافه شده در مقایسه با روش های وزن دهی معناگرای موجود، چندان بالا نمی باشد.

۲–۵– کارهای آینده

تعدادی کارهای آتی در مورد ادامه ی فعالیت بر روی وزن دهی معنایی و گرامری با دیدهای ارائه شده وجود دارد که در ادامه به آنها اشاره خواهد شد، تا مورد توجه و بررسی علاقمندان قرار گیرد.

یکی از چالش های پیش روی روش پیشنهادی وابستگی نظریه ی مرکزیت به زبان مرجع مجموعه ی اسناد است. یک معیار وزن دهی مانند TF-IDF به لحاظ وابستگی مطلق به ویژگی های آماری مانند فرکانس، هیچگونه وابستگی به زبان مرجع ندارد. اما تئوری های زبانی مانند نظریه ی مرکزیت و یا حتی حدس های شهودی که در

مورد اولویت نقش های گرامری ارائه می شود، می توانند باتوجه به زبان مرجع تغییر کنند. به صورت خاص قوانین ارائه شده برای نظریه ی مرکزیت در این پایان نامه در مورد زبان انگلیسی تحقیق و ثابت شده اند. نظریه ی مرکزیت در تعدادی دیگر از زبان ها که در قسمت مرور ادبیات از آنها نام برده شده است، نیز ارزیابی شده است. اما تاکنون ارزیابی از آن، به عنوان مثال برای زبان فارسی ارائه نشده است. لازمه ی استفاده از روش پیشنهادی در زبانی مانند فارسی، ارزیابی تئوری و یا تجربی نظریه ی مرکزیت بر روی زبان فارسی است. پس بسط روش پیشنهادی و تبدیل آن به یک روش چند زبانه از کارهای باز پیش رو می باشد.

یکی از کارهای پیش رو، بسط توانایی نظریه ی مرکزیت در کاربرد خاصی مانند وزن دهی به کلمات، به کل مجموعه ی اسناد می باشد. همانطور که گفتیم تنها توانایی نظریه مرکزیت در تشخیص انسجام و برجستگی محلی به اثبات رسیده است. شهودهای ارائه شده درباره اولویت انتقال های معنایی نسبت به هم، بر اساس همین اثبات ها بیان شده است. اما در مورد توانایی نظریه مرکزیت در تشخیص انسجام و برجستگی سراسری در مجموعه ای از اسناد، نتایج خوبی به دست نیامده است. هرچند که تاکنون تلاشهایی با استفاده از ابزارهای موجود در علوم کامپیوتر و ریاضی برای اثبات این فرضیه که نظریه مرکزیت قابلیت تشخیص انسجام و برجستگی سراسری را دارد و یا نه، صورت نپذیرفته است. یکی از کارهای پیش رو استفاده از یک روش یادگیری ماشینی برای تشخیص تعدادی از انتقال های معنایی که به صورت زنجیروار رخ می دهند، به عنوان زنجیره ی نشان دهنده ی انسجام و برجستگی سراسری است. البته این یادگیری باید در یک کاربرد خاص مانند خلاصه سازی چند سنده – برای انسجام محلی و تعیین اولویت محلی انتقال های معنایی از کاربرد آن در خلاصه سازی تک سنده استفاده گردید – صورت گیرد. به این وسیله می توان اولویت انتقال های معنایی و یا زنجیره های انتقال معنایی در مجموعه ای از اسناد را مشخص نمود. حال برای وزن دهی به انتقال های معنایی می توان یک گام جلوتر رفت و به وزن دهی سراسری به انتقال های معنایی و یا زنجیره ای از انتقال های معنایی پرداخت. البته در این گام نیز استفاده از روش های یادگیری ماشینی بسیار مثمر ثمر به نظر می آید.

در یکی از کارهای پیش رو می توان به پارامتریک کردن ضرایب مورد استفاده در این پایان نامه پرداخت. همانطور که دیدیم، هم برای وزن دهی به انتقال های معنایی و هم در وزن دهی به نقش های گرامری ضرایبی بر اساس مجموعه سند مورد پردازش مورد استفاده قرار گرفتند. به نظر می رسد استفاده از روش های محاسبات نرم مانند فازی در محاسبه و به کارگیری این ضرایب، می تواند به سازگار شدن روش پیشنهادی با مجموعه سند مورد استفاده و بهبود کارایی روش کمک زیادی نماید.

یکی از مشکلات پیش رو در بسیاری از روش های بازیابی اطلاعات حجم بالای ابعاد و فضای جستجو می باشد. یکی از کارهای مورد نظر، حذف تعدادی از کلمات و یا جملات بر اساس انتقال معنایی رخ داده میان جملات و یا نقش گرامری کلمات می باشد. به عنوان مثال زمانی که جمله ای درگیر انتقال معنایی Rough-shift می باشد، شاید بتوان از نقش گرامری مانند قید زمان در این جمله صرفنظر کرد. در انتخاب این کلمات و کاندید کردن آنها می توان از روش های یادگیری ماشینی نیز استفاده کرد.

منابع

۶- منابع

[AMI04] Amig´o, E., Gonzalo J., Peinado V., Pe˜nas, A., Verdejo, F. "Using syntactic information to extract relevant terms for multi-document summarization". Proceedings of the 20th international conference on Computational Linguistics (COLING04). 2004.

[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.

[BAR09] 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.

[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.

[BEN08] Bengtson, E., Roth, D. "Understanding the value of features for coreference resolution". Proceedings of the Conference on Empirical Methods in Natural Language Processing, October 25-27, Honolulu, Hawaii, 2008.

[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.

[BON02] Bontcheva, K., Cunningham, H., Tablan, V., Maynard, D., Hamza, O. "Using GATE as an Environment for Teaching NLP". Department of Computer Science University of Sheffield Sheffield, S1 4DP, UK, 2002.

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

[GAG05] Gagnon, M., Da Sylva, L. "Text summarization by sentence extraction and syntactic pruning". In Proceedings of Computational Linguistics in the North East (CliNE05), 2005.

[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.

[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.

[CHA97] Charniak, E. "Statistical Techniques for Natural Language Parsing". 1997.

[CHI99] Chisholm, E., Kolda, T.G. "New weighting formulas for the vector space method in Information Retrieval". March 1999 .

[COH95] Cohen, J. "Highlights: Language- and Domain-Independent Automatic Indexing Terms for Abstracting". Journal of the American Society for Information Science, volume 46(3), pp. 162-174, 1995.

[CRO94] Crochemore, M., Rytter, W. "Text Algorithms". Oxford University Press , pp.73-104, 1994.

[CUM08] Cummins, R. "The Evolution and Analysis of Term-Weighting Schemes in Information Retrieval". PhD thesis, National University of Ireland, Galway, 2008.

[DAS07] Das, D., Martins, A. "A Survey on Automatic Text Summarization". Literature Survey for the Language and Statistics II Course at CMU, 2007.

[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.

[DAT81] Date, C.J. "An introduction to database systems". Addison-Wesley, Reading, Mass, 1981.

[DEB04] D e b o l e, F., S e b a s t i a n i, F. "Supervised Term Weighting for Automated Text Categorization". In Text Mining and its Applications. S. Sirmakessis (Ed.), Heidelberg, Physica-Verlag, DE, pp. 81-98, 2004.

[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.

[EFR05] E f r o n, M. "Eigenvalue Based Model Selection during Latent Semantic Indexing". Journal of American Society for Information Science and Technology, Vol. 56(9), pp. 969-988, 2005.

[ELA05] Elango, P. "Coreference Resolution: A Survey". Technical Report, University of Wisconsin Madison, 2005.

[ERI99] E r i c, C., K o l d a, T.G. "New Term Weighting Formulas for the Vector Space Method in Information Retrieval". Technical Report, ORNL/TM-13756, Oak Ridge National Laboratory, 1999.

[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.

[FAL95] Falaoutsos, C., Oard, D.W. "A survey of information retrieval and ltering methods". Technical report CS-TR-3514, University of Maryland, College Park, MD20742, Aug. 1995.

[FUH91] Fuhr, N., Buckley, C. "A probabilistic learning approach for document indexing". ACM Transactions on Information Systems, volume 9(3), pp. 223–248, 1991.

[GIL02] Gildea, D., Jurafsky, D. "Automatic labeling of semantic roles". Computational Linguistics, volume 28(3), pp. 245-288, September 2002.

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

[GON01] Gong, Y., Liu, X. "Generic text summarization using relevance measure and latent semantic analysis". Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR`01, New Orleans, Louisiana,United States, 2001.

[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.

[GRE98] Greiff, W. "A theory of term weighting based on exploratory data anal-ysis". In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, Mel-bourne, Australia, 1998.

[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.

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

[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.

[HOF96] Hoffman, B. "Summarization: an Application for NL Generation". Proceeding of International workshop on NL Generation,1996.

[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.

[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.

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

[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.

[KRO93] Krovetz, R."Viewing morphology as an inference process". In R. Korfhage et al., Proc. 16th ACM SIGIR Conference, Pittsburgh, pp. 191-202, June 27-July 1, 1993.

[KUM08] K u m a r, C h. A s w a n i, S r i n i v a s, S. "Identifying the Number of Dimensions for Dimensionality Reduction in Latent Semantic Indexing". In Proc. of 2nd International Conference on Information Processing. Bangalore, India, pp. 64-70, 2008.

[KUM09] K u m a r, C h. A s w a n i, S r i n i v a s, S. "On the Performance of Latent Semantic Indexing Based Information Retrieval". Journal of Computing and Information Technology, 2009.

[LEE95] Lee, J.H. "Combining multiple evidence from different properties of weighting schemes". Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 180-188, 1995.

[LEE97] Lee, D.L., Chuang, H. "Document Ranking and the Vector-Space Model", IEEE Software, volume 14(2), pp. 67-75, 1997.

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

[LUH57] Luhn, H. P. "Statistical approach to mechanized encoding and searching of literary information". IBM Journal of Research and Development, volume 1(4), pp. 309–317, 1957.

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

[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.

[MOE00] Moens, M.F. "Automatic indexing and abstracting of document texts". Kluwer Academic Publishers, 2000.

[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.

[POR80] Porter, M.F. “An algorithm for suffix stripping”, Program, volume 14(3), pp. 130-137, 1980.

[RAD02] Radev, D. R., Hovy, E., McKeown, K. "Introduction to the special issue on summarization". Computational Linguistics., volume 28(4), pp. 399-408, 2002.

[REI85] Reichman, R. "Getting Computers to Talk Like You and Me". The MIT Press, Cambridge, MA, 1985.

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

[ROB76] Robertson, S.E., Sparck Jones, K. "Relevance weighting of search terms". Journal of the American Society for Information Science, volume 27, pp. 129-146, 1976.

[ROC66] Rocchio, J.J. "Document Retrieval Systems - Optimization and Evaluation". PhD thesis, Harvard, 1966.

[SAL75] Salton, G., Wong, A., Yang, C.S. "A vector space model for automatic indexing". Commun ACM, volume 18(11), pp. 613–620, 1975.

[SAL83] Salton, G., Fox, E.A., Wu, H. "Extended boolean information retrieval".Commun ACM, volume 26(11), pp. 1022–1036, 1983.

[SAL88] Salton, G., Buckley, C. "Term-weighting approaches in automatic text Retrieval". Information Processing & Management, volume 24(5), pp. 513–523, 1988.

[SAL89] Salton, G. "Automatic text processing: The transformation, analysis, and retrieval of information by computer". Addison-Wesley, Reading, MA, 1989.

[SAN05] Sanderson, M., Zobel, J. "Information retrieval system evaluation: effort, sensitivity, and reliability". In: SIGIR ’05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, Salvador, Brazil, pp. 162–169, 2005.

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

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

[SID04] Siddharthan, A., Nenkova, A., McKeown, K. "Syntactic Simplification for Improving Content Selection in Multi-Document Summarization". In Proceedings of COLING, 200.

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

[SPA72] Sparck Jones, K. "A statistical interpretation of term specificity and its application in retrieval". Journal of Documentation, volume 28(1), pp. 11–21, 1972.

[SPA73] Sparck Jones, K. "Index term weighting". Information Storage and Retrieval, volume 9, pp. 619–633, 1973.

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

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

[SWI04] Swier, R., Stevenson, S. "Unsupervised semantic role labeling". Proceedings of the 2004 Conference on Emperical Methods in Natural Language Processing, Barcelona, Spain, pp. 95–102, 2004.

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

[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.

[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.

[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.

[YU82] Yu, C.T., Lam, K., Salton, G. "Term weighting in information retrieval using the term precision model". Journal of the ACM (JACM), volume 29(1), pp. 152–170, 1982.

[ZHU06] Z h u, M., G h o d s I, A. "Automatic Dimensionality Selection from the Screen Plot via the Use of Profile Likelihood". Computational Statistics and Data Analysis, volume 51(2), pp. 918-930, 2006.

[کام۸۹] کامیار، حسین.، کاهانی، محسن.، کامیار، محسن.، پورمعصومی حسن کیاده، آصف. " روش جدید خلاصه سازی استخراجی تک سندی با استفاده از نظریه مرکزیت ". شانزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، ایران، تهران، صفحه ۴۷۹-۴۸۸، اسفند ۱۳۸۹.

ضمائم

۷– ضمائم

۱–۷– ضمیمه الف– مراحل پنج گانه یک سیستم بازیابی اطلاعات

۱–۱–۷– ایندکس بندی

در طی ایندکس بندی، سندها برای استفاده توسط سیستمِ بازیابی اطلاعات آماده می شوند. این به معنی تبـدیل و آماده سازی مجموعهٔ خام از اسناد به نمایش قابل دسترس و ساده از اسناد است. این تبدیل- از متن سند بـه یـک نمایش از متن – به عنوان ایندکس بندی شناخته می شود.

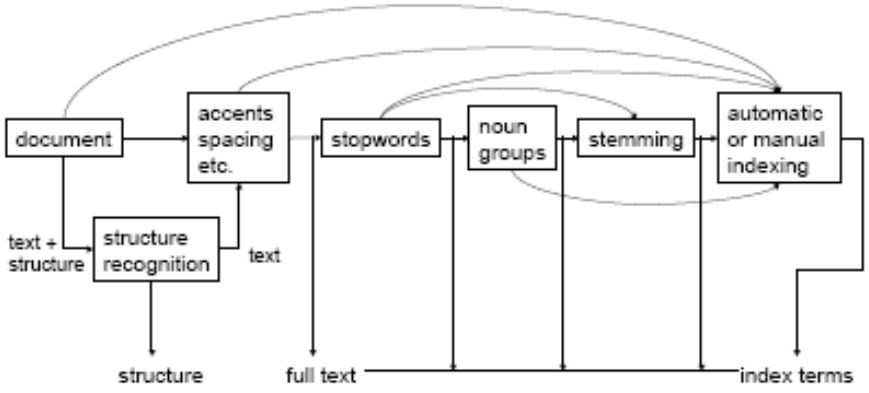
تبدیل یک سند به فرم ایندکس بندی شده شامل استفاده از

۱. یک کتابخانه یا مجموعه ای از عبارات منظم
۲. پارسرها
۳. یک کتابخانه از کلمات ایست و ربط (اضافه)[1]
۴. فیلترهای متفرقهٔ دیگر

این اعمال بطور طبیعی در ۵ مرحله صورت میگیرد: حذف فرمت و قالب و نشانه هـا، تـوکن بنـدی، فیلتـر کـردن، ریشه یابی و وزن گذاری. اگر هیچ گونه حذف نشانه های زبان های میانی مانند HTML و وزن گذاری نیـاز نباشـد آنگاه این تبدیل تنها شامل تشخیص توکن ها، فیلتر کردن و ریشه یابی می شود. این نوع از ایندکس بندی متناوباً در پایگاه داده هایی که صرفاً فایلهای متنی و اطلاعات خام را مرتب می کنند یافت می شود، با ایـن وجـود در وِب بدلیل اینکه سند ها در فرمتهای مختلف هستند پنج مرحلهٔ فوق نیاز است [COH95].

شمای کلی از تبدیل یک متن به کلمات ایندکس شده در شکل ۷-۱ دیده می شود.

[1] - StopList, Stop word such: If,a,Then,For,…



شکل ۷-۱- مراحل تبدیل یک متن به کلمات ایندکس شده

۲-۱-۷- خطی سازی سند ها

خطی سازی سند عملی است که به موجب آن یک سند به رشته ای از عبارتها (terms) تبدیل می گردد [CRO94]. همانطور که در ادامه می آید این عمل در دو مرحله انجام می گیرد:

۲. **حذف قالب ها و نشانه ها:** در طی این فاز تمامی تگ ها، نشانه گذاریها و قالب بندیهای خاص از سند حذف می شوند. بنابراین برای یک سندِHTML تمامی تگها و متن های داخل این تگ ها حذف می گردند. این عمل به طور طبیعی شامل تمامی صفات عناصر، اسکریپت ها، خطوط توضیحی و متون قرار داده شده داخل آنها می باشد. بعضی از موتورهای جستجوی تجاری ممکن است متن داخل تگ عنوان، image ALT attribute، صفات خلاصهٔ جدول(table summary attribute) و تگِ فرا توضیح (Meta Desciption tag) را حفظ کنند. دیگرِ سیستمها ممکن است از صفات عناصر یا متادِیتاها اصلاً محافظتی نکنند و حتی بعضی دیگر ممکن است اینها را نگهداری کنند و با اطلاعات متادیتاها کاربر نهائی را کمک کنند، اما این اطلاعات را برای رتبه بندی اسناد مورد استفاده قرار ندهند [CRO94].

۳. **توکن بندی (Tokenization):** در طی این فاز بقیهٔ متن پارس(تجزیه) میشود و تمامی حروف کوچک و بزرگ به یک نوع تبدیل شده و نقطه گذاریها حذف می شوند.قوانین هایفن[۲]گذاری بایستی اعمال شوند، بطور مثال در بعضی سیستمها ممکن است این هایفن ها حفظ شوند در حالیکه در بقیه

[1]- parse

[۲]- Hyphen: خط ربط برای پیوستن هجاهای جدا شده از یکدیگر

ممکن است هایفن ها حذف شده و یا به عنوان فضای خالی و یا توکنِ ربط در نظر گرفته شوند.بعضی از موتورهای جسنجو مانند گوگِل، در سمت کوئری به نظر می رسد که هایفن ها را به عنوان قسمتی از جستجوی دقیق محلی (localized EXACT[1] search) در نظر می گیرند(به طور مثال در حالت FINDALL[2] در کوئری به فرمِ K1-K2 + K3، بخشِ K1-K2 از این کوئری به عنوان یک کوئریِ دقیق محلی (localized EXACT query) تفسیر می شود [CRO94].

در طیِ خطی سازی تمامیِ دستورالعمل های CSS (Cascading Style Sheet) حذف می شوند، این بدان معنی است که بدون داشتن فهمی روشن از پروسهٔ خطی سازی متن، انتقال های دلخواه متن با استفاده از CSS، تگ های تودرتو یا جداول، فی الواقع می تواند مضر باشد بطوریکه آنچه که کاربران به عنوان محتوای مربوط درک می کنند با آنچه که موتور جستجو به عنوان مربوط می خواند و نمره دهی می کند تفاوت پیدا میکند. در واقع پس از خطی سازی سند:

۱. جریان مؤثر متن بایستی دنبالهٔ منسجمی از لغات را داشته باشد.
۲. این دنباله متن بایستی معانی، تِم ها، موضوعات و زیر فهرستهای مورد نظر متن اصلی را شامل باشد.
۳. موقعیت لغات در این دنباله متنی توسط خطوطِ Markup که در Source Code واقع است مشخص می گردد (مانند HTML Tags).

تمامی اینها این واقعیت را تاکید می کند که درک کاربر از مرتبط بودن[3] ودرک ماشین از مرتبط بودن دو چیز متفاوت است. اغلب اوقات روشی که موتورهای جستجو یک متن را درک و تفسیر می کنند با روشی که کاربران و مرورگرها[4] آن متن را می خوانند تفاوت دارد و این دلیل آن است که خطی سازی سند – به عنوان بخشی از آنالیز فاصله (GAP analysis) – قبل و بعد از بهینه سازی سند اهمیت دارد. در برخی موارد، خطی سازی سند تلاش برای محاسبهٔ چگالی کلمه بوسیلهٔ اسکنِ یک سند را بیهوده می کند [CRO94].

۳-۱-۷- فیلتر کردن

فیلترکردن به عمل تصمیم گیری برای اینکه کدامین کلمات (terms) بایستی برای نمایش یک سند انتخاب شوند، اطلاق می گردد که این میتواند برای موارد زیر مورد استفاده قرار گیرد:

---

[1]- جستجو در حالت Exact باعث می شود که موتور جستجو سندهایی را برگرداند که ترتیب کلمات همانند ترتیب کلمات کوئری باشد.

[2]- جستجو در حالت Findall تنها منجر به بازگرداندن سندهایی با ترتیبِ کلماتی همانند ترتیب کلماتِ کوئری نمی شود، بلکه سندهایی با ترتیب دلخواه از کلمات کوئری را نیز شامل می شود.

[3]- relevancy

[4]- browser

۱. توضیح متن سند

۲. تشخیص یک سند از اسناد دیگر در مجموعه ای از اسناد.

خیلی از اوقات کلمات با استعمال زیاد، به دو علت نمی تواند به عنوان انتخاب مورد استفاده قرار گیرند. اول آنکه تعداد سندهایی که به کوئری مربوط هستند نسبت کوچکی از مجموعه اسناد هستند. کلمه ای که در جداسازی اسناد مربوطه از اسناد نامربوط مؤثر است، کلمه ای است که به احتمال زیاد در تعداد کمی از اسناد آمده است؛ این بدان معنی است که کلماتی که تعداد تکرارشان زیاد است برای تشخیص سند ضعیف هستند. دلیل دوم آن است که کلماتی که در متن های مختلفی تکرار می شوند، عنوان یا زیر عنوان یک سند را نمی توانند تعریف کنند [SAL89].

به همین دلایل است که کلمات با استعمال زیاد و یا کلماتِ ایست(Stop Word) از رشتهٔ کلمات حذف می شوند، اگرچه حذف لغات ایست[1]از یک سند کاری وقت گیر است. یک روش بهینه آن است که تمامیِ کلماتی که بطور مشترک در مجموعهٔ اسناد ظاهر شده اند و تاثیری بر بازیابی اسناد مربوط ندارند، حذف شوند.

این کار می تواند با داشتن کتابخانه ای از کلمات ایست- لیست کلمات ایست (Stop List) برای حذف – انجام گیرد. این لیست ها (از کلمات ایست) می تواند عمومی (اعمال شونده به تمامی مجموعه اسناد) ویا مشخص (ساخته شده برای مجموعه اسناد مشخص) صورت گیرد. اینکه چه حدی از تکرار کلمات مشخص می کند که کدام کلمات بایستی از مجموعه اسناد حذف شوند، وابسته به پیاده سازیهای مختلف است. بطور مثال در بعضی از سیستمهای بازیابی اطلاعات کلماتی که در بیش از ۵۰٪ از مجموعه اسناد ظاهر شده اندف حذف می شوند. در بعضی از سیستمها کلماتی که در لیست کلمات ایست نیستند اما در بیش از ۵۰٪ از مجموعه اسناد آمده اند به عنوان " کلمات منفی" (Negative Terms) در نظر گرفته می شوند و آنها نیز بخاطر جلوگیری از پیچیدگی در وزن دهی حذف می شوند. این مورد اخیر در بعضی از سیستمهای مبتنی بر MySQL که از وِزن هایِIDFP (Probabilistic Inverse Document Frequency) استفاده میکنند، استفاده می گردد [SAL89].

۴–۱–۷– ریشه یابی

ریشه یابی چون یکی از مراحل اصلی پیش پردازشی الگوریتم پیشنهادی این پایان نامه بوده است، در اصل پایان نامه مورد بررسی قرار گرفته است.

[1]- Stop Word

۲-۷- ضمیمه ب- انواع مدل های نمایش مجموعه اسناد

۱-۲-۷- مدل شمارِ کلمات[1]

همانطور که در قبل اشاره شد، وزن کلمات با استفاده از اطلا عات سراسری و محلی محاسبه می شود. در مدل فضای برداری کلاسیک وزن کلمهٔ i اُم با استفاده از تساوی زیر محاسبه می شود:

$$w_i = tf_i * \log\left(\frac{D}{df_i}\right)$$

در این قسمت بحثی غیر تکنیکی بر روی تئوری بردار کلمه داریم.

۱-۱-۲-۷- شمارِ کلمات

همانطور که اشاره شد، در حالیکه رویکرد استاندارد برای بسیاری از سیستمهای بازیابی اطلاعات همان وزنهای tf*IDF است، بصورت تاریخی وزنها با استفاده از اطلاعات محلی محاسبه می شوند [YU82]:

وزن کلمه = $w_i = tf_i$ (۷-۱)

در این مدل و با هر الگوی وزن دهی دیگر، به سه عنصر احتیاج است [GRE98]:

- یک پایگاه داده تا اسناد را از آن بازیابی کنیم.
- یک کوئری(پرس و جو) به عنوان ورودی.
- ایندکسی از کلمات.

دو عامل اول بصورت آشکار مشخص است.*ایندکس کلمه*[2]ٔخزنی از کلمات یکتاست که بر طبق قوانین انتخابِ مشخص (Selection Rules) از اسناد استخراج شده اند.

۲-۱-۲-۷- محاسبهٔ بزرگی بردارها

طبق تعریف؛ هر برداری دارای جهت و بزرگی است، حال این کمیتها را محاسبه می کنیم.

---

[1] - Term Counts Model
[2]- Index Term

برای محاسبهٔ بزرگی هر بردار قصیهٔ فیثاغورس را بکار می بریم: برای n بُعد می توان نوشت [YU82]:

$$|D_i| = \left(a_1^2 + a_2^2 + a_3^2 + ... + a_n^2\right)^{1/2}.$$

۳-۱-۲-۷- محاسبهٔ *حاصلضرب داخلی*(اسکالر)

قبل از محاسبهٔ جهت بردارها (مقدار کسینوس) نیازمند دانستن حاصلضرب داخلیِ بردار کوئری و بردار سند هستیم.

حاصلضرب نقطه ای، حاصلجمعِ حاصلضرب های شمارِ کلمات و شمارِ کوئری است.

نکته ای در حاصلضرب داخلی: بعضی از حاصلضرب داخلی به عنوان معیاری برای شباهت و نزدیکی سند و کوئری استفاده می کنند.محاسبات به صورت عملیات ماتریسی می تواند در آید که درآن حاصلضرب داخلی (اسکالِر) به عنوان معیار شباهت در نظر گرفته می شود [GRE98].

$$Sim(Q, Dj) = Q \bullet Dj = q^T * Dj \quad (۷-۲)$$

که q^T ماتریس ترانهادهٔ کوئری است.

آشکارا مشخص است که سند هایی که شامل تمامی کلمات (terms) کوئری باشند، بیشترین نمرهٔ شباهت را به خود اختصاص می دهد.

ماتریس کلمه-سند A به صورت زیر تعریف می شود:

A=[D1,D2,D3,...,Dn]

نمرات وابستگی برای کوئریِ Q بوسیلهٔ ضرب ماتریسیِ $q^T * A$ بیان می شود که نتیجهٔ آن برداری n بُعدی است که نمرهٔ وابستگیِ هر سند را نشان می دهد.

۴-۱-۲-۷- محاسبهٔ جهت ها (مقدار کسینوسی)

---

[۱]- حاصلضرب نقطه ای:Dot Product

اکنون باید مقدار متناظر کسینوسی محاسبه شود، به دلیل اینکه حاصلضرب داخلی بصورت حاصلضرب بزرگی بردارها در کسینوس زاویهٔ بین آنها تعریف می شود. به زبان ساده، بردار کوئری را در فضای برداری سندها منعکس کرده ایم و کسینوس زاویهٔ بین بردار کوئری وبردار هر سند در مجموعهٔ اسناد موجود را محاسبه کرده ایم. ایده کلی آن است که بررسی کنیم کدامین سند به بردار کوئری نزدیکتر است (با زاویهٔ کمتر) [YU82].

۵-۱-۲-۷- رتبه بندی نتایج

حال بایستی اسناد را بر طبق مقدار کسینوس رتبه بندی و مرتب کنیم. در حالت کلی آن سندی که مقدار کسینوسش نزدیک به یک باشد آن سند مرتبط تر است و اگر مقدار کسینوسش صفر شود آنگاه سند و کوئری در فضای کلمه ای[1]برهم عمودند و این بدان معنی است که سند و کوئری به هم مربوط نیستند(نمونه سند۳) [GRE98].

۶-۱-۲-۷- محدودیت های مدل

مدل *شمار کلمات* (Term Counts) [FAL95] :

۱. حساس به تکرار کلمات استLee,Chuang وSeaman شش مدل وزن دهی کلمهٔ متفاوت از تساوی۲-۹استخراج کرده اند[2]و دریافته اند که هیچ وضعیتی پیدا نمی شود که درآن مدلی بهتر عمل کند [LEE97].
۲. در سند های بزرگ از آنجائیکه این اسناد شامل تعداد زیادی کلمه هستند که اغلباً تکرار شده اند و ماتریس کلمه- سند آنها دارای عناصر زیادی است بنابراین این اسناد دارای نمرهٔ شباهت بیشتری نسبت به اسناد کوچکتر هستند بدون آنکه واقعاً مربوط تر باشند.

عملیات بهتر آن است که با در نظر گرفتن اطلاعات محلی و سراسری، tf را در IDF ضرب کنیم. بدین طریق می توان حجم اطلاعاتی که به کلماتِ پرس و جو شده حساس است را داخل کرد. بنابراین در محاسبات قبلی نیاز است که $w_i = tf_i$ را با $w_i = tf_i * IDF$ جایگزین کرده و ماتریس کلمه-سند را با این مقادیر پر کنیم.

۲-۲-۷- مدل فضای برداری کلاسیک

[1]- Term Space
[2]- Document Ranking and the Vector-Space model

بر عکس مدل شمارِکلمات، مدل فضای برداری سالتون اطلاعات محلی و سراسری را با هم ترکیب می کند $w_i = tf_i * \log\left(\frac{D}{df_i}\right)$. نسبت df_i / D احتمال انتخاب سندی ازبین مجموعهٔ اسناد است که شامل کلمهٔ جستجو شده (queried) باشد [SAL75]. این می تواند به عنوان احتمال سراسری در تمام مجموعه باشد. بنابراین عبارت $\log\left(\frac{D}{df_i}\right)$ فرکانس سند معکوس (IDFi) و نمایندهٔ اطلاعات سراسری است.شکل زیر رابطهٔ بین فرکانس محلی و سراسری را در یک پایگاه دادهٔ ایده آل که شامل پنج سند D1 تا D5 است را نشان می دهد. تنها سه سند شامل عبارت "car" می باشند. اگر از این سیستم برای این کلمه کوئری بگیریم مقدار IDF برابر log5/3=0.2218 بدست می آید.

۱-۲-۲-۷- المان های خود متشابه

المان های خود متشابه یکی از ویژگی های هندسهٔ فراکتالی است. طبیعت این اشکال شامل خود-متشابهی آنها چندین درجه تشخیص می باشد. هر مجموعه شامل اسناد، هر سند شامل عبارات[1] هر عبارت شامل جملات وکلمات می باشد.بنابراین برای هر کلمهٔi در سندj می توانیم از فرکانس کلمه در مجموعه (Cf)، فرکانس کلمه (tf)، فرکانس عبارت (Pf)، و فرکانس جمله (Sf) صحبت کنیم [SAL75].

$$\mathrm{Cf}_{\mathrm{i}} = \sum_{j} \mathrm{tf}_{\mathrm{i,j}} \qquad ((\text{الف})۷-۳)$$

$$\mathrm{Pf}_{\mathrm{i,j,p}} = \sum_{\mathrm{P}} \mathrm{Pf}_{\mathrm{i,j,p}} \qquad ((\text{ب})۷-۳)$$

$$\mathrm{Pf}_{\mathrm{i,j,p}} = \sum_{\mathrm{s}} \mathrm{Sf}_{\mathrm{i,j,p,s}} \qquad ((\text{ج})۷-۳)$$

تساوی ۷-۳(ب) تلویحاً تساوی۲-۹ را اشاره می کند.مدلهایی که تلاش می کنند تا وزن کلمات را با مقادیر فرکانس بیان می کنند، بایستی میزان مرتبط بودن را در نظر بگیرند. یقیناً اصطلاح "چگالی کلمهٔ کلیدی" که توسط بهینه سازان موتورهای جستجو (SEO) ترویج داده می شود، در این طبقه بندی نیستند.

۲-۲-۲-۷- محدودیت های مدل

[1]- Passages

به عنوان یک مدل پایه، الگوی بردار کلمه محدودیت های زیادی را دارد؛ اول آنکه دارای محاسبات زیادی است. از لحاظ محاسباتی این مدل بسیار کُند است و به زمان زیادی نیاز دارد، ثانیاً اگر یک کلمهٔ جدید به فضای کلمات وارد شود نیاز داریم تا تمامی بردارها دوباره محاسبه شوند. همانطور که چونگ و سیمن اشاره کرده اند [LEE97]، محاسبهٔ طول بردار کوئری(اولین عبارت در مخرج تساوی ۷-۴) نیاز به دسترسی به نه تنها کلمات مشخص در کوئری بلکه به تمام کلمات سند نیز دارد.

بقیهٔ محدودیت ها عبارتند از:

۵. اسناد بزرگ: محاسبهٔ مقادیر تشابه برای سندهای بزرگ مشکل است.
۶. تطابق های منفی اشتباه[1]:اسناد با محتوای مشابه ولی واژگان متفاوت ممکن است منجر به حاصلضرب های داخلی کوچکی شود. این محدودیت های سیستم های بازیابی اطلاعات بر پایهٔ کلمه می باشد.
۷. تطابق های مثبت اشتباه[2]:جمله بندیهای نادرست، حذف پیشوند و پسوند و تجزیه می تواند منجر به hitهای جعلی شود (کارگر، کار+گر ؛therapist,the + rapist) این تنها یک محدودیت ماقبل پردازش است نه اینکه محدودیت مدل برداری باشد.
۸. محتوای معنایی: سیستمهایی که برای اداره کردن محتوای معنایی بکار می روند ممکن است نیاز به استفاده از تَگ های خاصی داشته باشد (containers).

می توان مدل را با استفاده از راهکارهای زیر بهتر کرد

۱. تهیه کردن مجموعه ای از کلمات کلیدی که مشخصهٔ هر سند باشد.
۲. حذف تمامی کلمات ایست و کلمات متداول ("از"، " به"، "برای"، "for"، "then")
۳. ریشه یابی کلمات به ریشه های آنها.
۴. محدود کردن فضای برداری به اسمها و فعلها و صفات توصیفیِ محدود.
۵. استفاده از فایلهای امضاء (دیجیتال) و یا استفاده از فایلهای معکوس (Inverted Files) نه چندان بزرگ.
۶. استفاده از تکنیکهای نگاشت موضوعی.
۷. محاسبهٔ زیر بردارها یا passage vectors در اسناد بزرگ.
۸. بازیابی نکردن اسنادِ زیرِ آستانهٔ مقدار مشابهتِ تعریف شدهٔ کسینوسی.

[1]- False Negative Matches
[2]- False Positive Matches

۳-۲-۷- مدلهای برداری با فرکانسهای نرمال شده

۱-۳-۲-۷- وزن کلمات وKeyword Spamming

در مدلهای شمارِ کلمات و مدل کلاسیک بردارِ کلمه ای داریم:

مدل شمارِ کلمات: $w_i = tf_i$

مدلِ کلاسیک: $w_i = tf_i * \log\left(\frac{D}{df_i}\right)$

وزن wi از کلمهٔ i متناسب با فرکانس آن یعنی tfi در نظر گرفته می شود، از آنجائیکه کلمات با تکرار بیشتر وزن بیشتری از کلمات با تکرار کمتر به خود می گیرندبنابراین هر دو مدل نسبت به هرزنامه نویسی کلمات کلیدی[1] آسیب پذیرند. هرزنامه نویسی کلمات کلیدی روشی مخرب است که در آن کلمات به قصد بالابردن رتبهٔ سند در موتورهای جستجو، از روی عمد تکرار می شوند. در این روند رتبه دهی و بازیابی به خطر می افتد. این مدل های بردار کلمه ای با نرمال کردن سند و کوئری کمتر مستعد خطا توسط هرزنامه نویسی کلمات کلیدی هستند.

۲-۳-۲-۷- فرکانس های نرمال شده سند

فرکانس نرمال شدهٔ کلمهٔ i در سندِ j با تساوی زیر بیان می گردد [SAL88].

$$f_{i,j} = tf_{i,j} / \max tf_{i,j} \quad (۷-۴)$$

بطوریکه $f_{i,j}$ فرکانس نرمال شده، $tf_{i,j}$ فرکانس کلمهi در سند j، $\max tf_{i,j}$ ماکزیمم فرکانس کلمهٔ i در سند jمی باشند.

بطور مثال ؛ سندِ شامل تعداد کلماتِ زیر را در نظر بگیرید:

Major: ۱ بار baseball:۴ بار league:۲ بار playoffs:۵ بار

از آنجائیکه"playoffs" بیشترین تکرار را داشته بنابراین فرکانس های نرمال شده برابر است با:

[1] - Keyword spamming

"league":۲/۵=۰,۴۰ "playoffs":۵/۵=۱ "major":۱/۵=۰,۲۰ "baseball":۴/۵=۰,۸۰

۳-۳-۲-۷- فرکانس های نرمال شده کوئری

فرکانس نرمال شدهٔ کلمهٔ i در کوئری Q با تساوی زیر بیان می شود [SAL88].

$$f_{Q,i} = 0.5 + 0.5 * tf_{Q,i} / \max tf_{Q,i} \quad (۷-۵)$$

بطوریکه $f_{Q,i}$ فرکانس نرمال شده، $tf_{Q,i}$ فرکانس کلمهi در کوئری Q و $\max tf_{Q,i}$ ماکزیمم فرکانس کلمهٔ i در کوئریQ .

بطور مثال برای کوئریِ"major major league"=Q فرکانس ها برابر است با:

Major: ۲ بار league: ۱ بار

از آنجائیکه "major" بیشترین رخداد را داشته است، بنابراین فرکانس های نرمال شده برابر است با:

"major":(۰,۵ + ۰,۵ ×۲/۲)=۱ "league": (۰,۵ + ۰,۵ ×۱/۲)=۰,۷۵

۴-۳-۲-۷- وزن های نرمال شده

وزن کلمهi در سند j بصورت زیر می تواند نوشته شود [SAL88]:

$$w_{i,j} = \frac{tf_{i,j}}{\max tf_{i,j}} * \log\left(\frac{D}{df_i}\right) \quad (۷-۶)$$

و وزن کلمهٔ i در کوئری Q بصورت زیر نوشته می شود [SAL88]:

$$w_{Q,i} = \left(0.5 + 0.5 * \frac{tf_{Q,i}}{\max tf_{Q,i}}\right) + \log\left(\frac{D}{df_i}\right) \quad (۷-۷)$$

باید توجه داشت که هرکسی با تساویهای۲-۱۶ و۲-۱۷ موافق نیست.ضمناً فرکانس هـای نرمـال شـده مقـدار نـیم (۰,۵) آزاد می گیرند(حتی برای $tf_{Q,i}$=۰).

۴-۲-۷- مدل گلاسکو [1]

در طول سالیان گذشته طرح های وزن دهی نرمال شدهٔ مختلفی ارائه شده است، یک طرح جاالب توسط ساندرسون[2] در یک گزارش ارائه شده است. آنها عبارت زیر را پیشنهاد می کنند [SAN05]:

$$w_{ij} = \frac{\log\left(freq_{ij}+1\right)}{\log\left(length_j\right)} \bullet \log\left(\frac{N}{n_i}\right) \quad (۷-۸)$$

بطوریکه $w_{i,j} = tf * idf$ وزن کلمهٔ i در سند j، $freq_{ij}$ فرکانس کلمهٔ i در سند j، $length_j$ تعداد کلمات یکتا در سند j، N تعداد اسناد موجود در مجموعه و n_i تعداد اسنادی که کلمه i در آن ظاهر شده است، می باشند.

این طرح توسط عوامل زیر بر کوئری اعمال می شود:

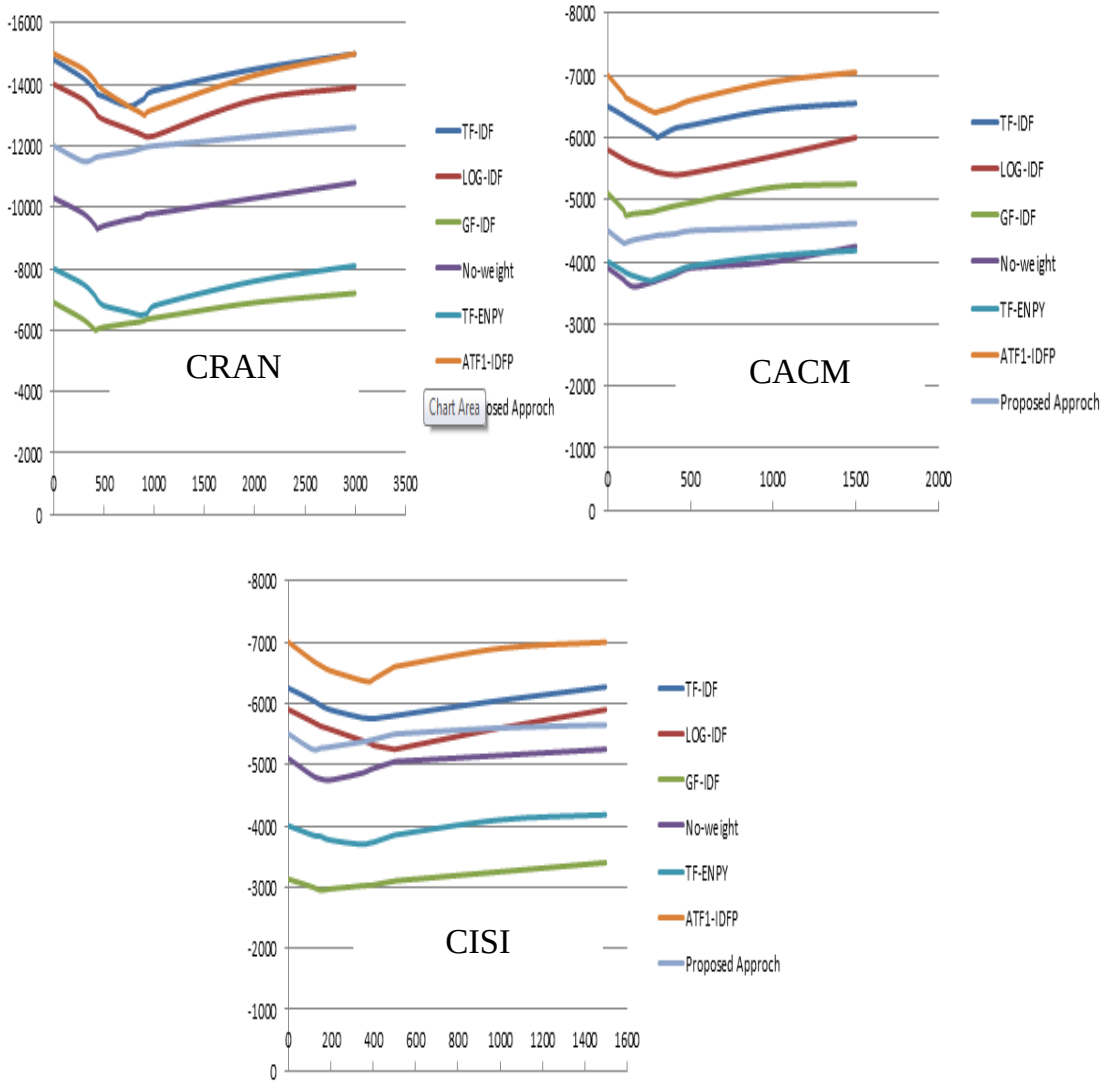
۱. استفاده از فرکانس های نرمال شده و تعریف شده در تساوی ۷-۸.
۲. تعریف تعداد کلماتِ شامل کلمات ایست و ربط به عنوان طول اسناد و کوئری ها.

باید توجه داشت که تمامی مدلهایی که بررسی شد همگی از طرح وزن دهی سراسریِ یکسان که به معنی همان IDF هاست، استفاده می کنند.

۳-۷-ضمیمه ج- اشکال نتایج آزمون های انتخاب ابعاد حقیقی

۱-۳-۷- آزمون انتخاب ابعاد حقیقی، تابع PL

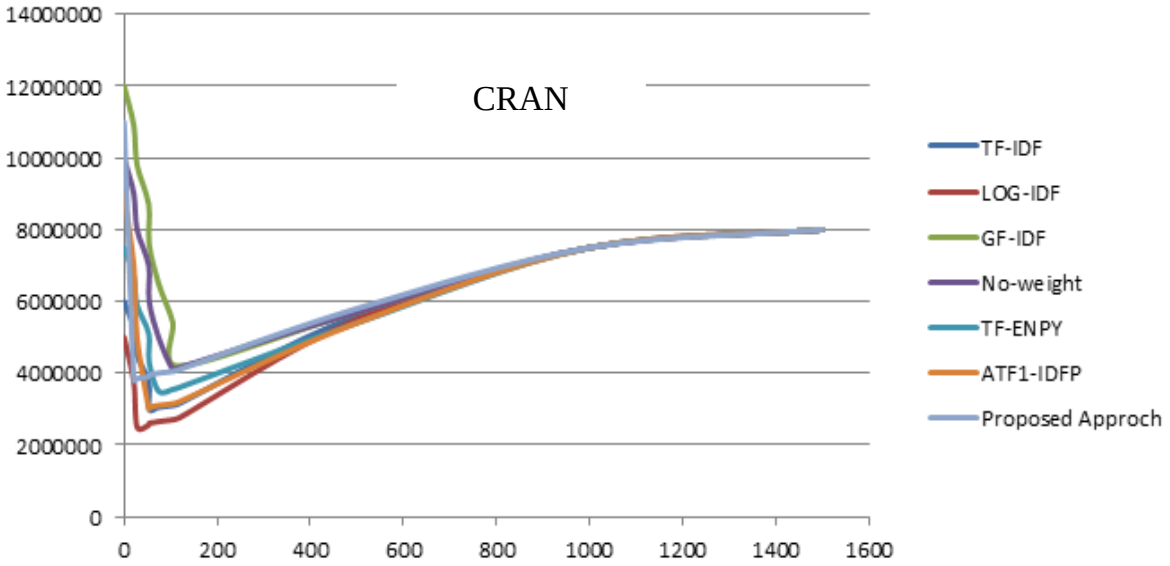
[1]- Glasgow Model
[2]- Mark Sanderson

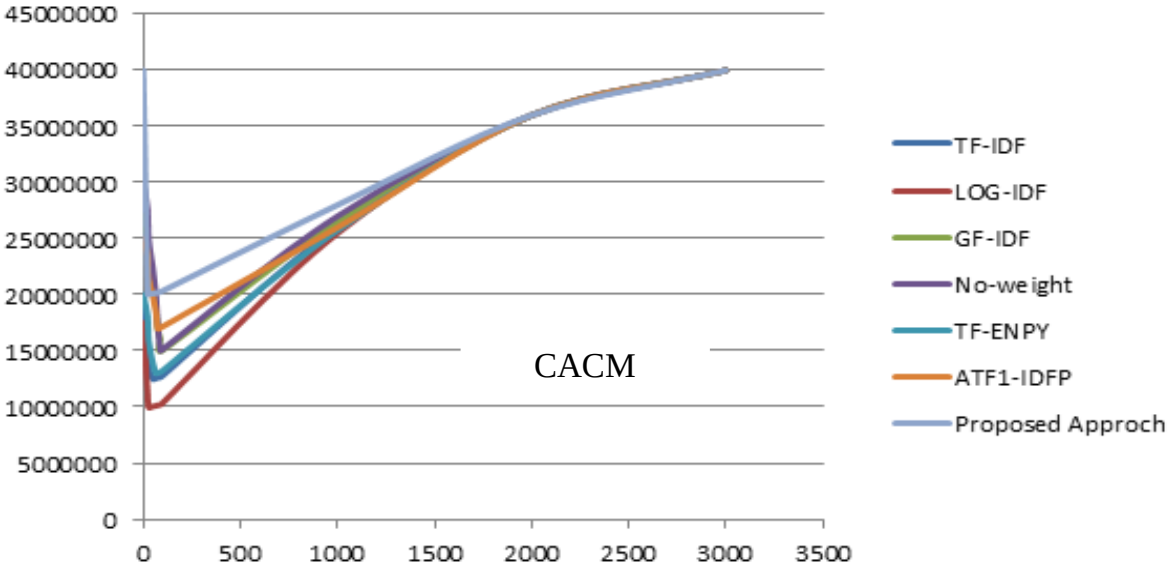


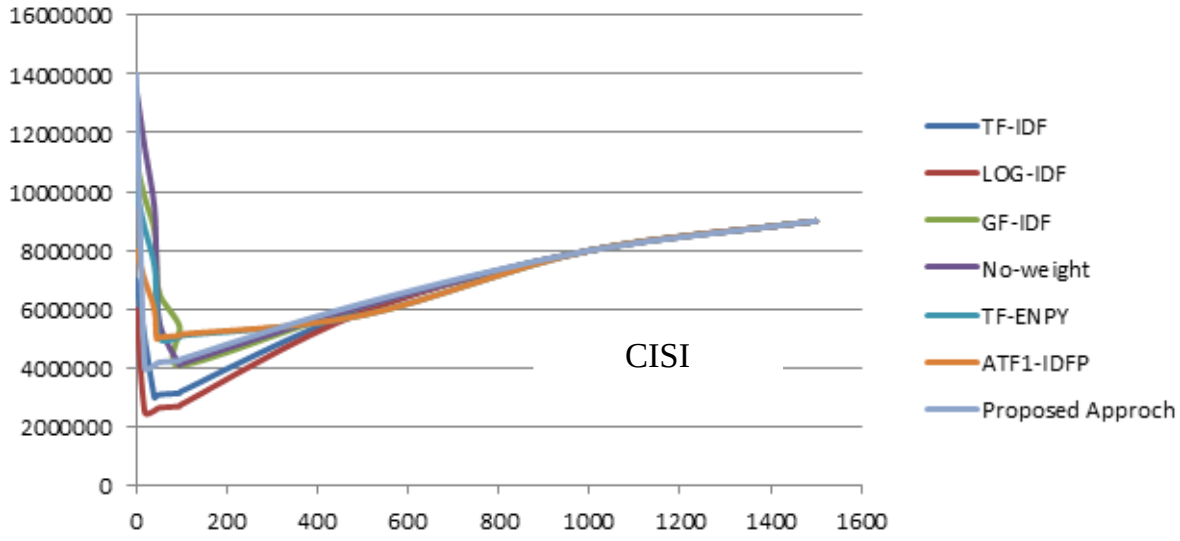
محور x ها: تعداد ابعاد (Number of Dimensions) و محور y ها: مقادیر شباهت همسایگی (Likelihood Value)

شکل ۷-۲- نمودارهای مقادیر ابعاد حقیقی بازگردانده شده برای مجموعه داده های سه گانه به ازای معیارهای وزن دهی مختلف توسط تابع PL

۲-۳-۷- آزمون انتخاب ابعاد حقیقی، تابع AMDL


CRAN
14000000
12000000
10000000
8000000
6000000
4000000
2000000
0
0
200
400
600
800
1000
1200
1400
1600
TF-IDF
LOG-IDF
GF-IDF
No-weight
TF-ENPY
ATF1-IDFP
Proposed Approch

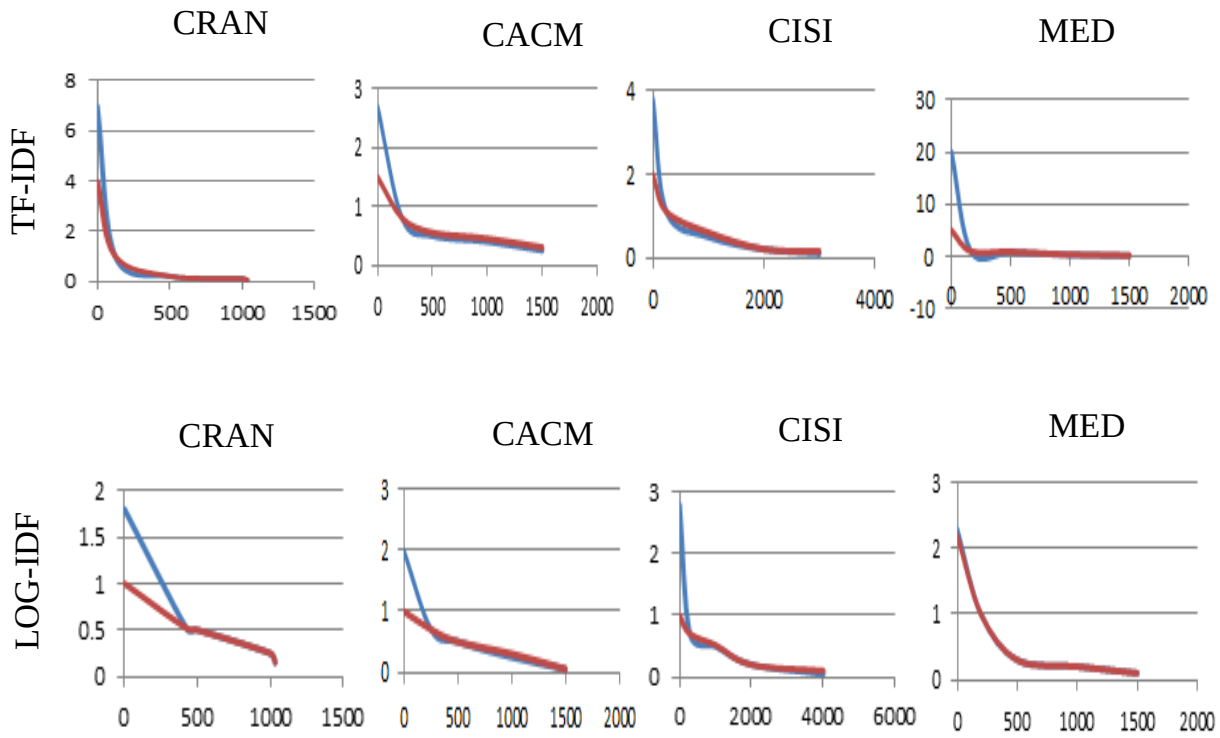

CACM
45000000
40000000
35000000
30000000
25000000
20000000
15000000
10000000
5000000
0
0
500
1000
1500
2000
2500
3000
3500
TF-IDF
LOG-IDF
GF-IDF
No-weight
TF-ENPY
ATF1-IDFP
Proposed Approch

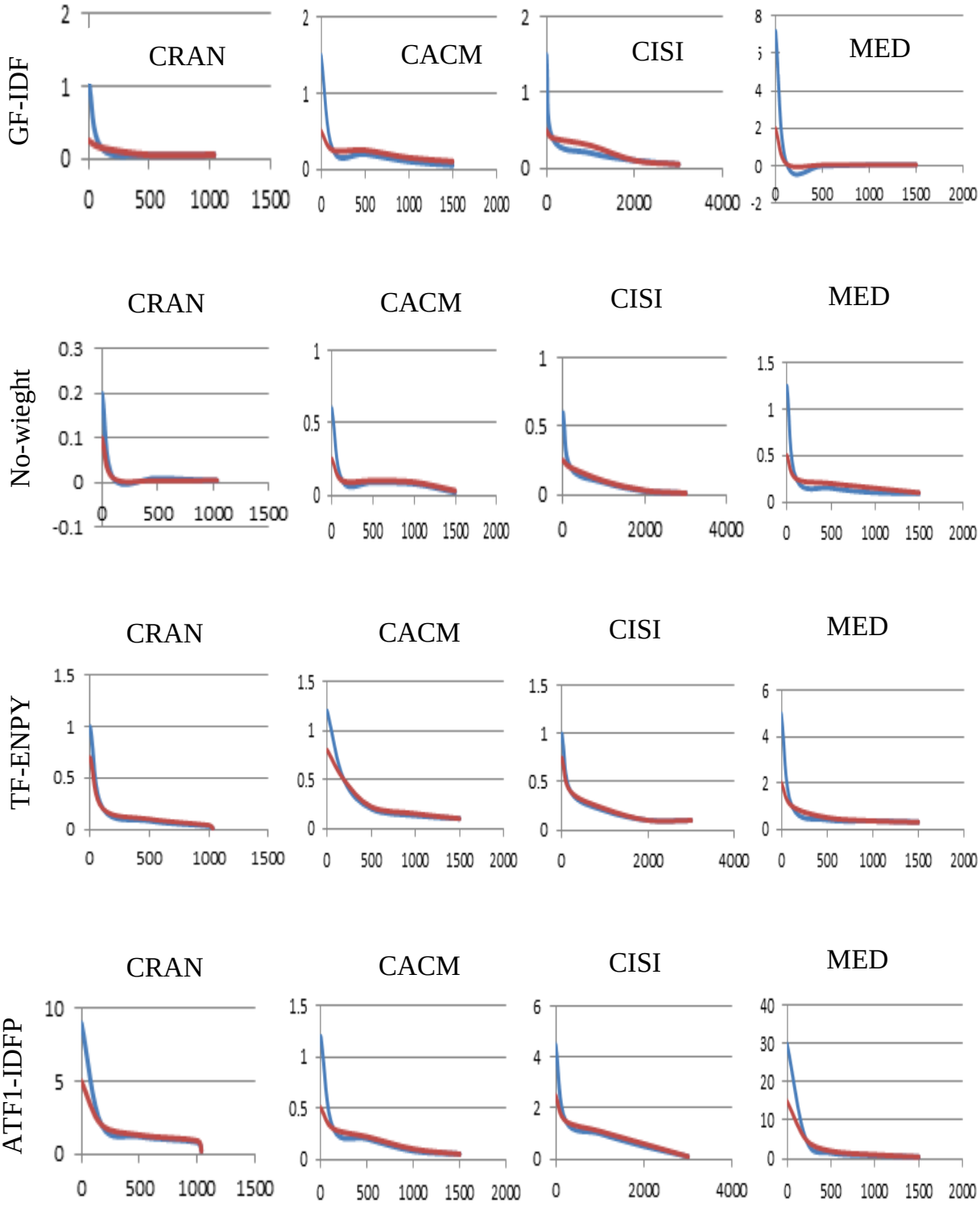


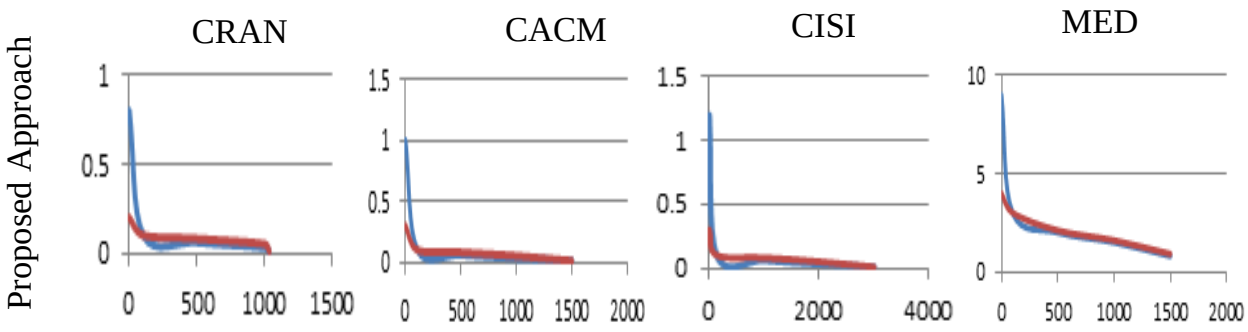
محور x ها: تعداد ابعاد (Number of Dimensions) و محور y ها: مقادیر تابع AMDL

شکل ۷-۳- نمودارهای مقادیر ابعاد حقیقی بازگردانده شده برای مجموعه داده های سه گانه به ازای معیارهای وزن دهی مختلف توسط تابع AMDL

۳-۳-۷- آزمون انتخاب ابعاد حقیقی، تابع PA




GF-IDF
CRAN
CACM
CISI
MED
No-wieght
CRAN
CACM
CISI
MED
TF-ENPY
CRAN
CACM
CISI
MED
ATF1-IDFP
CRAN
CACM
CISI
MED



محور x ها: تعداد ابعاد (Number of Dimensions) و محور y ها: مقادیر ویژه (Eigenvalues)

شکل ۷-۴- نمودارهای نقطه برخورد ماتریس های تصادفی و واقعی مقادیر ویژه برای مجموعه داده های چهارگانه به ازای معیارهای وزن دهی مختلف توسط تابع PA