
بسم الله الرحمن الرحيم

دانشکده: کامپیوتر و فناوری اطلاعات

رشته: مهندسی کامپیوتر گرایش: هوش مصنوعی و رباتیک

پایان‌نامه کارشناسی ارشد

تعریف یک تابع فاصله فازی جهت مقایسه عبارات

نگارنده:

مهرداد کاهه

استاد راهنما:

دکتر مرتضی زاهدی

شهریور ۱۳۹۶

فرم شماره ۷: صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با تأییدات خداوند متعال و با استعانت از حضرت ولی عصر (عج) ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد آقای **مهرداد کاهه** به شماره دانشجویی **۹۳۱۳۹۶۴** رشته **مهندسی کامپیوتر** گرایش **هوش مصنوعی** تحت عنوان **تعریف یک تابع فاصله فازی جهت مقایسه عبارات** که در تاریخ **۹۶/۰۶/۰۸** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می‌گردد:

قبول (با درجه : خوب امتیاز ۱۶) ☐ دفاع مجدد ☐ مردود ☐

نوع تحقیق: نظری ☐ عملی ☐

۱ـ عالی (۲۰ ـ ۱۹)
۲ـ بسیار خوب (۱۸/۹۹ ـ ۱۸)
۳ـ خوب (۱۷/۹۹ ـ۱۶)
۴ـ قابل قبول (۱۵/۹۹ ـ ۱۴)
۵- نمره کمتر از ۱۴ غیر قابل قبول

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول	دکتر مرتضی زاهدی		
۲- استادراهنمای دوم	ـــــ		
۳- استاد مشاور	ـــــ		
۴- نماینده شورای تحصیلات تکمیلی	دکتر منصور فاتح		
۵- استاد ممتحن اول	دکتر محمد قاسم زاده		
۶ـ استاد ممتحن دوم	دکتر محسن رضوانی		

نام و نام خانوادگی رئیس دانشکده: دکتر علی‌اکبر پویان

تاریخ و امضاء و مهر دانشکده:

تقدیم به پدر و مادر عزیز و مهربان و صبورم

به خواهر دوست داشتنی ام

و تمامی خوبان زندگی که درس ایستادگی و محبت آموختند.

تشکر و قدردانی

از استاد راهنمای عزیز و بزرگوارم، جناب آقای دکتر مرتضی زاهدی که علاوه بر راهنمایی های بی دریغشان،

استادی دلسوز و گرانقدر در زندگی ایجانب هستند، کمال تشکر و قدردانی را دارم.

تعهد نامه

اینجانب **مهرداد کاهه** دانشجوی دوره کارشناسی ارشد رشته هوش مصنوعی دانشکده کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **تعریف یک تابع فاصله فازی جهت مقایسه عبارات** تحت راهنمائی آقای دکتر مرتضی زاهدی متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام‌شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورداستفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود است و مقالات مستخرج با نام «دانشگاه صنعتی شاهرود» و یا «Shahrood University of Technology» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافتهای آن‌ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است. **تاریخ**

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود است. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان‌نامه بدون ذکر مرجع مجاز نمی‌باشد.

چکیده

ساخت سیستمی که بتواند به صورت موثری فاصله بین دو عبارت را تعیین کند همواره مورد توجه محققین زمینه‌های هوش مصنوعی و بازیابی اطلاعات بوده است. معیارهای شباهت دو عبارت در طیف گسترده‌ای از برنامه‌های کاربردی همچون پردازش زبان‌های طبیعی، اصلاح پرس‌وجوهای جستجو، غلط‌یابی معنایی، جلوگیری از دستبردهای علمی، مقایسه اسناد و دیگر زمینه‌های کاربردی در بازیابی اطلاعات به کار می‌رود. ورودی سیستم‌های تشخیص‌دهنده می‌تواند یک متن و یا مجموعه‌ای از اسناد و متون باشد و خروجی آن نتیجه قضاوت سیستم در مورد مشابهت جملات ورودی متون و میزان شباهت آنها است. در نهایت نزدیک بودن قضاوت سیستم به قضاوت انسان نشان‌دهنده عملکرد خوب سیستم است.

در این تحقیق روشی برای محاسبه شباهت لغوی مبتنی بر عدم قطعیت ارائه می‌نماییم. برای این منظور از روش‌های مبتنی بر اثر انگشت و تکنیک Winnowing به عنوان تکنیک‌های محاسبه شباهت استفاده می‌نماییم. همچنین با در نظر گرفتن چندین پارامتر جهت محاسبه شابهت عبارات، از سیستم استنتاج فازی برای تصمیم‌گیری در شرایط عدم قطعیت استفاده نموده‌ایم. در این روش ویژگی‌هایی مبتنی بر تکنیک هش‌گذاری و دگردیسی‌های کلی جملات تعریف شده است که این معیارها از سرعت و دقت مناسبی برخوردار هستند. پس از محاسبه شباهت متون، جملات مشابه نیز توسط سیستم استخراج می‌گردد. روش مذکور بر روی پیکره PAN فارسی و پیکره ارزیابی شباهت عبارات دانشگاه صنعتی شاهرود مورد ارزیابی قرار گرفته است که با دقت ۷۸ درصد، اسناد را به سه کلاس متفاوت، نسبتاً مشابه و مشابه تفکیک می‌نماید.

کلمات کلیدی: شباهت سنجی لغوی، سرقت ادبی، سیستم استنتاج فازی، اثر انگشت، winnowing

فهرست مطالب

فصل اول: مقدمه

فصل دوم: مروری بر روش‌های محاسبه شباهت متون

فصل سوم: روش پیشنهادی

فصل چهارم: جزئیات پیکربندی و ارزیابی روش پیشنهادی

فصل پنجم: بحث و نتیجه گیری

فهرست اشکال

فهرست جداول

فصل اول

مقدمه

۱- مقدمه

۱-۱- انگیزش و هدف پژوهش

امروزه با توجه با افزایش حجم اطلاعات و مستندات در زمینه‌های مختلف، دسترسی سریع به اطلاعات مورد نظر برای هر فرد از اهمیت ویژه‌ای برخوردار است. از این رو علاوه بر تکنیک‌های بازیابی اطلاعات، تکنیک‌های خلاصه‌سازی و دسته‌بندی می‌تواند کمک موثری در افزایش سرعت دسترسی کاربران به اسناد مورد نظرشان باشد. ساخت سیستمی که بتواند به صورت موثری فاصله بین دو عبارت را تعیین کند همواره مورد توجه بسیاری از محققین در زمینه‌های هوش مصنوعی و بازیابی اطلاعات بوده است. تعیین فاصله بین دو عبارت می‌تواند از طریق محاسبات آماری و اندازه‌گیری شباهت بین واژه‌ها و یا توسط رویکردهای مبتنی بر یادگیری ماشین صورت پذیرد. معیارهای شباهت دو عبارت در طیف گسترده‌ای از برنامه‌های کاربردی همچون پردازش زبان‌های طبیعی، اصلاح پرس‌وجوهای جستجو، غلط‌یابی معنایی، جلوگیری از دستبردهای علمی، مقایسه اسناد و دیگر زمینه‌های کاربردی در بازیابی اطلاعات به کار می‌رود. در مقایسه با کامپیوتر انسان‌ها قادر به اندازه‌گیری شباهت دو عبارت با دقت بسیار زیاد هستند، که این بر اساس حس و دانشی است که در آنها وجود دارد. در تکنیک‌های بررسی شباهت میان دو عبارت، با جایگزینی کلمات مترادف، عبارت مشابه و جملات هم‌معنا در متن و یا با جایگزینی الگوها انجام می‌پذیرد. ورودی سیستم‌های تشخیص‌دهنده می‌تواند یک کلمه، پاراگراف و یا مجموعه‌ای از اسناد و متون باشد و خروجی آن نتیجه قضاوت سیستم در مورد مشابه بودن یا نبودن جملات ورودی متون و میزان شباهتشان است. در نهایت نزدیک بودن قضاوت سیستم به قضاوت انسان نشان‌دهنده عملکرد خوب سیستم است.

محاسبه شباهت بین متون و عبارات از رویکردهای متفاوتی موردبررسی قرار می‌گیرد. رویکرد مبتنی بر روش‌های آماری، رویکرد مبتنی بر به کارگیری الگوریتم‌های یادگیری ماشین، استفاده از درخت‌های تصمیم، رویکردهای مبتنی بر تعداد کلمات کلیدی مشترک بین جمله‌ها، رویکرد مبتنی بر شباهت نحوی، شبکه‌های عصبی راهکارهای متفاوتی را در مقایسه دو عبارت ارائه می‌کنند.

از آنجا که ذهن انسان در تشخیص شباهت‌ها عملکرد بسیار خوبی دارد پس ما نیز سعی می‌کنیم تا حدودی راه و روش انسان را اقتباس نماییم. انسان‌ها در تشخیص شباهت‌ها از عباراتی نظیر کم، زیاد و نسبتا شبیه برای ارزش به جمله‌ها استفاده می‌کنند و با استفاده از روابط موجود در حافظه این بررسی‌ها را انجام می‌دهند. برای این منظور می‌توان از یک سیستم استنتاج فازی بهره گرفت. به این صورت که این سیستم استنتاج با استفاده از مجموعه‌ای از پارامترها مانند کلمات کلیدی مشترک بین عبارت‌ها، تعداد کلمات یکسان و تعداد جملات در نهایت نتایج حاصل از استنتاج به صورت فازی را به یک مقدار قطعی تبدیل نماید. طبیعتاً هر چه این مقدار بیشتر باشد نشان دهنده این امر است که دو عبارت ارتباط نزدیک‌تری با یکدیگر دارند.

۱-۲- تعریف مسئله

با توجه به آنچه که گفته شد، محاسبه شباهت بین دو سند در بیشتر کارهای پردازش متن یک عملیات کلیدی به شمار می‌رود. دو متن با درصد بالایی از اشتراک کلمات و عبارات را مشابه به یکدیگر می‌گویند. معیارهای متفاوتی برای یافتن شباهت متون از جمله معیارهای مبتنی بر رشته[۱]، معیارهای شباهت معنایی[۲] و معیارهای مبتنی بر پیکره[۳] تعریف شده است [۱] که در بخش‌های آتی به تفصیل شرح داده می‌شوند. ما

---

۱ **String Base Similarity**

۲ **Semantic Base Similarity**

۳ **Corpus Base Similarity**

در پژوهش پیش‌رو قصد داریم تا با استفاده از معیارهای مبتنی بر رشته و الگوریتم‌های مبتنی بر اثر انگشت[۱] و هش‌گذاری و همچنین تعریف معیارهای جدید میزان تشابه اسناد را مورد سنجش قرار دهیم. از آنجا که شباهت متون در شرایط عدم قطعیت صورت می‌پذیرد، سیستم استنتاج فازی[۲] را که همانند انسان بر اساس داده‌های ورودی تصمیماتی اتخاذ می‌کند، به عنوان موتور تصمیم‌گیری شباهت دو سند بر اساس سری ویژگی‌های تعریف شده توسعه داده‌ایم. این سیستم با دریافت مجموعه قواعد فازی که به عنوان ویژگی‌های سند به سیستم داده می‌شود شباهت دو سند را ارزیابی می‌کند. در نهایت با وافازی‌گری[۳] خروجی فازی، مشابهت دو سند با عددی بین صفر و یک نشان داده می‌شود که هر چه شباهت دو سند بیشتر باشد این مقدار به یک نزدیک‌تر خواهد بود [۲].

لازم است ابتدا تعریفی بر مبنای ورودی‌های سیستم داده ارائه گردد. در این سامانه، ورودی‌ها به سه گروه کم، میانه و زیاد دسته‌بندی می‌شوند. ورودی‌های کم شامل کلمات، عبارات و جملات کوتاه هستند. ورودی‌های میانه شامل متون با چند پاراگراف است. ورودی‌های با مقیاس زیاد شامل اسناد طولانی بیش از هزار کلمه است. نحوه دسته‌بندی ورودی‌ها به این سه دسته در فصل سوم تشریح می‌شود. در نهایت با در نظر گرفتن این ورودی‌ها از الگوریتم‌های متفاوتی برای محاسبه شباهت استفاده خواهیم کرد.

پس از تقسیم‌بندی عبارات ورودی بر اساس مقیاس ورودی‌ها آنها شباهت آنها محاسبه می‌شود. در ورودی‌های با مقیاس میانه و زیاد، از الگوریتم‌های تغییریافته مبتنی بر اثر انگشت استفاده می‌شود. در این روش از ورودی‌ها n-گرام تهییه می‌شود و پس از استخراج n-گرام‌ها، طبق فرآیند معروف به انگشت‌نگاری به اثر انگشت و یا شینگل‌ها تبدیل می‌شوند. در این مرحله هش‌های ورودی‌ها با یکدیگر مقایسه شده و بر

---

[۱] **Fingerprint**

[۲] **Fuzzy Inference System**

[۳] **Defuzzification**

اساس معیار همپوشانی[۱] شباهت بین آنها محاسبه می‌گردد. همچنین دو معیار شباهت دگردیسی کلی[۲] و دگردیسی محلی[۳] نیز تعریف می‌گردد. معیار دگردیسی کلی بر پایه محاسبه فرکانس کاراکترهای بین دو تکه از متن شباهت بین ورودی‌ها را محاسبه می‌نماید. معیار دگردیسی محلی بر اساس جابجایی، حذف و جایگزینی کاراکترهای درون بخش محدودتری از متن در حد عبارات و گرام‌ها، به محاسبه شباهت می‌پردازد. در نهایت با تعریف قواعد شباهت توسط متخصصین و اعمال آنها به سیستم استنتاج فازی[۴]، خروجی سیستم استنتاج فازی مشخص‌کننده شباهت بین دو ورودی خواهد بود. لازم به ذکر است که خروجی سیستم استنتاج فازی در سه دسته متفاوت، نسبتاً مشابه و مشابه قرار می‌گیرد که در بخش‌های بعدی مفصلا شرح داده خواهد شد.

جهت ارزیابی سامانه ارائه شده، از پایگاه داده PAN استفاده نموده‌ایم [۳]. این پایگاه داده در سال ۲۰۱۵ توسط خوشنوا و همکاران با اقتباس از اسناد فارسی موجود در پیکره[۵] ویکی‌پدیا صورت گرفته است که مشتمل بر حدود سه هزار جفت سند با مقیاس‌های متفاوت می‌باشد که در برخی از آنها سرقت علمی-ادبی[۶] رخ داده است. بر اساس آنچه که مدنظر ما می‌باشد، شابهت جفت ورودی‌های این پیکره طی فرآیندی به سه دسته متفاوت، نسبتاً مشابه و مشابه تقسیم‌بندی شده اند که در فصل چهارم به بیان چگونگی آن خواهیم پرداخت. در انتها نحوه عملکرد سیستم در تشخیص شابهت اسناد با سایر تحقیقات صورت گرفته مقایسه می‌شود.

---

[۱] **Overlap**

[۲] **Global Deformation**

[۳] **Local Deformation**

[۴] **Fuzzy Inference System**

[۵] **Corpus**

[۶] **Plagiarism**

۱–۳- کاربردهای محاسبه شباهت بین اسناد

از شباهت‌یابی بین اسناد در زمینه‌های گوناگونی همچون شباهت‌یابی دو فایل صوتی، تصویری و متنی استفاده می‌شود. اما آنچه که بنا داریم در این تحقیق در مورد آن به بحث و پژوهش بپردازیم شباهت‌یابی متون با تفکیک آنها به سه دسته کم، میانه و زیاد است. از کاربردهای محاسبه شباهت می‌توان به موارد زیر اشاره کرد:

۱. اصلاح غلط‌های تایپی و نگارشی.
۲. شناسایی سرقت علمی-ادبی و حفظ حق نشر.
۳. تشخیص شباهت کدهای نرم‌افزاری.
۴. استخراج پاسخ‌های مشابه در سیستم‌های پرسش و پاسخ تعاملی.
۵. دسته‌بندی اسناد.
۶. یافتن عبارات در موتورهای جستجو [۴].

۱–۴- ساختار پایان‌نامه

در فصل دوم، ابتدا به بررسی الگوریتم‌های مرتبط ارائه شده در حوزه شباهت متون از سه دیدگاه مبتنی بر رشته، مبتنی بر پیکره و مبتنی بر پایگاه دانش می‌پردازیم. پس از آن برخی از روش‌های رایج مبتنی بر رشته و محتوا در محاسبه شباهت متون و تشخیص سرقت علمی-ادبی را بررسی خواهیم نمود. همچنین در این بخش به تشریح سیستم استنتاج فازی خواهیم پرداخت. در ادامه برخی از پژوهش‌های مرتبط انجام شده داخلی و خارجی مرتبط با روش پیشنهادی را ارائه خواهیم کرد. در انتها برخی از چالش‌های زبان فارسی مرتبط با موضوع تحقیق را بیان می‌کنیم که در روش پیشنهادی سعی در بهبود آنها داریم.

در فصل سوم به توسعه سیستم استنتاج فازی محاسبه شباهت خواهیم پرداخت. در این فصل، در ابتدا معماری کلی سیستم معرفی خواهد شد. پس از آن تک تک اجزای سیستم را که شامل سه بخش پیش‌پردازش، استخراج ویژگی و استنتاج فازی است تشریح خواهیم نمود. در این بخش روشی را ارائه می‌نماییم که علاوه بر محاسبه شباهت بین متون، جملات کاندید شباهت را نیز استخراج کند و به عنوان بخش‌های مشابه نمایش دهد.

در فصل چهارم، در ابتدا به معرفی پایگاه داده ارزیابی سیستم می‌پردازیم. همچنین برخی تغییراتی که نیاز است برای ارزیابی سیستم در پایگاه داده اعمال شود نیز به تفصیل شرح داده می‌شود. همچنین پایگاه داده ارزیابی شباهت بین عبارات آزمایشگاه وب‌کاوی دانشگاه شاهرود، که برای ارزیابی این تحقیق تکمیل و گردآوری شد نیز معرفی می‌گردد. در نهایت خروجی حاصل از اعمال داده‌های آزمایشی مورد ارزیابی قرار می‌گیرد و سیستم از نظر کارایی با سیستم مشابه مبتنی بر گراف مورد مقایسه قرار خواهد گرفت.

در فصل پنجم نیز خلاصه‌ای از روش پیشنهادی و نتایج را ارائه خواهیم نمود. همچنین پیشنهاداتی را برای بهبود نتایج سیستم مد نظر قرار می‌دهیم که می‌تواند ادامه تحقیق ما و سایر پژوهشگران حوزه متن‌کاوی قرار گیرد.

فصل دوم

مروری بر معیارهای محاسبه شباهت متون

۲- مروری بر روش‌های محاسبه شباهت متون

۲-۱- بررسی شباهت از دیدگاه‌های متفاوت

محاسبه شباهت میان متون، نقشی اساسی در تحقیقات در زمینه پردازش متن ایفا می‌کند. این زمینه کاربردهای فراوانی در بخش‌های مختلفی از قبیل بازیابی اطلاعات، دسته‌بندی اسناد، انتخاب عنوان برای متن، تولید خودکار سوال، سیستم‌های خودکار پرسش و پاسخ، درجه‌بندی مقالات، نمره‌دهی سیستم‌های پاسخ کوتاه، ترجمه ماشینی و خلاصه متون دارد. در حوزه شباهت بین متون، در ابتدا مسئله محاسبه شباهت میان کلمات مطرح می‌گردد. در مراحل بعدی پس از محاسبه شباهت میان کلمات، شباهت میان جملات، پاراگراف‌ها و در نهایت متون بلند مطرح می‌گردد.

مشابهت کلمات می‌تواند از دو دیدگاه مختلف بررسی گردد. مشابهت می‌تواند خاص حرف و یا لغوی[۱] شود. در این حیطه ما تنها به ظاهر کلمات استناد می‌نماییم و معنا را دخیل نمی‌کنیم. اگر کلمات دارای رشته کاراکترهای مشابه باشند ما آنها را مشابه در نظر خواهیم گرفت. در غیر این صورت میان آنها شباهتی قائل نمی‌شویم [۱].

حیطه دیگر مورد بررسی، حیطه معنایی می‌باشد. در این حوزه به معنای درونی دو کلمه نیز توجه می‌کنیم. نمونه‌ای از شباهت‌های معنایی شامل مترادف بودن، کلمات متضاد و کلمات دارای طرح موضوع یکسان می‌باشند. به طور کلی هر دو دیدگاه محاسبه شباهت از دیدگاه حرف یا ظاهر و محاسبه شباهت از دیدگاه

[۱] **Lexical**

معنایی در سه دسته الگوریتم گوناگون مورد بررسی و ارزیابی قرار می گیرند. این سه دسته الگوریتم شامل روش‌های مبتنی بر رشته[۱]، روش‌های مبتنی بر پیکره[۲] و روش‌های مبتنی بر دانش[۳] هستند که حوزه لغوی در دسته مبتنی بر رشته مورد بررسی قرار می‌گیرد، و حوزه معنایی[۴] در دسته مبتنی بر پیکره و مبتنی بر دانش ارزیابی می‌شوند [۵].

معیارهای اندازه‌گیری مبتنی بر رشته با سری از کاراکترها و ترکیب کاراکترهای درون یک کلمه طراحی و پیاده‌سازی می‌شوند. معیارهای شباهت براساس پیکره، معیار محاسبه شباهت معنایی می‌باشند که شباهت بین کلمات و عبارات را طبق اطلاعات کسب شده از یک پیکره بزرگ[۵] محاسبه می‌کنند. معیارهای شباهت مبتنی بر دانش نیز از معیارهای محاسبه شباهت معنایی می‌باشند که درجه شباهت معنایی بیان کلمات را با استفاده از اطلاعات کسب شده از شبکه های معنایی[۶] ایجاد می‌کنند [۱].

در ادامه مروری کلی بر معروف‌ترین و پرکاربردترین الگوریتم‌هایی که در این سه حوزه معرفی شده‌اند خواهیم داشت.

۲-۱-۱- معیارهای شباهت مبتنی بر رشته

به طور کلی معیارهای شباهت مبتنی بر رشته بر روی سری قرارگیری رشته‌ها و ترکیب کاراکترها عملیات شباهت را انجام می‌دهند و در نهایت به یک معیار و درجه برای بیان شباهت میان دو ورودی می‌رسند.

---

۱ **String-based**

۲ **Corpus-based**

۳ **knowledge-based**

۴ **semantic**

۵ **corpora**

۶ **semantic networks**

درجه‌ای که ما از آن سخن می گوییم، متریک رشته[1] نام گذاری شده است. این متریک میزان شباهت یا عدم شباهت (فاصله) بین دو متن ورودی را محاسبه می‌کنند و مقدار فاصله را با عددی میان صفر و یک که بیانگر میزان شباهت است، نمایش می‌دهند.

الگوریتم‌های مشابهت مبتنی بر رشته، از دو جهت معیارهای مبتنی بر کاراکتر و معیارهای مبتنی بر عبارت مورد بررسی قرار می‌گیرند که در این بخش به هر یک از این معیارها و الگوریتم‌های مشمول آنها خواهیم پرداخت.

۲-۱-۱-۱- معیارهای شباهت مبتنی بر کاراکتر

در این بخش برخی از الگوریتم‌هایی که صرفاً وظیفه محاسبه شباهت دو عبارت یا رشته ورودی بر اساس کاراکترهای آنها را دارند، مورد بررسی قرار می‌دهیم. این معیارها شباهت بین دو رشته را بر اساس افزودن، حذف و یا تغییر کاراکترها تشخیص می‌دهند. از این الگوریتم‌ها معمولا برای تشخیص خطاهای تایپی استفاده می‌شود. یک معیار گویای مبتنی بر کاراکتر، حداقل تعداد عملیات ویرایش مورد نیاز برای تبدیل یک رشته به رشته‌ای دیگر را به روش‌های گوناگون ارزیابی می‌کند. عملیات‌های مجاز برای تبدیل یک رشته به رشته دیگر شامل قرار دادن، حذف، و جایگزینی هستند.

به عنوان مثال دو رشته‌ی زیر را در نظر بگیرید:

$S_1 = \text{“vldb”}$

$S_2 = \text{“pvldb”}$

$ED(\text{“vldb”}, \text{“pvldb”}) = 1$

[1] **string metric**

فاصله ویرایش[۱] عبارت S_1 و S_2 برابر ۱ است، زیرا عبارت اول را می‌توان با قرار دادن یک کاراکتر $''p''$ به دومی تبدیل کرد. دو رشته با توجه به معیار فاصله ویرایش در صورتی با هم مشابه هستند که فاصله ویرایش آنها بزرگتر از آستانه t نباشد. در ادامه به بررسی جزء جزء این الگوریتم‌ها می‌پردازیم.

۱. الگوریتم LCS[۲]

این الگوریتم روشی است که برای پیدا کردن بزرگترین زیردنباله در مجوعه‌ای از دنباله‌ها (غالبا دو دنباله) به کار می‌رود و جزء قدیمی‌ترین مسئله‌های علوم کامپیوتر است. این الگوریتم اساس کار برنامه‌های مقایسه‌کننده فایل‌ها می‌باشد که در نهایت تفاوت دو فایل را نمایش می‌دهد.

هدف مسئاله مقایسه این دو رشته و پیدا کردن شباهت بین آنها است. بزرگترین زیردنباله مشترک این طور تعریف می‌شود که دنباله‌ای مانند S_3 است به طوری که حروف موجود در S_3 با حفظ ترتیب در S_1 و S_2 موجود باشد. اما مطلقاً لزومی ندارد که متوالی باشد. از طرفی S_3 باید بزرگ‌ترین دنباله ممکن با خواص بالا باشد. قابل ذکر است که بلندترین زیر دنباله مشترک را می‌توان با استفاده از برنامه نویسی پویا پیدا کرد.

۲. الگوریتم Damerau- Leveneshtein:

این الگوریتم تابع فاصله بین دو رشته را به واسطه شمردن کم‌ترین تعداد عملیاتی که نیاز است تا یک رشته را به رشته دیگر تبدیل کند، به عنوان خروجی تعیین می‌کند. عملیات تعریف شده عبارت‌اند از

---

[۱] **Edit Distance**

[۲] **Longest Common Substring**

افزودن[۱]، حذف[۲]، جایگزینی[۳] که بر روی یک کاراکتر تعریف می‌شوند. عملیاتی دیگر نیز از قبیل ترانهاده[۴] کردن دو کاراکتر مجاور نیز گاها بر روی این عملیات تعریف می‌شوند [۶].

۳. الگوریتم جارو[۵]:

این الگوریتم بر اساس تعداد و محل قرارگیری کاراکترهای مشترک بین دو رشته معیار فاصله را ارزیابی می‌کند. مدل کامل شده آن جارو-وینکلر[۶] نام دارد. این الگوریتم از یک درجه پیشوند[۷] که یک رتبه بهتر به رشته‌هایی که از ابتدا با یکدیگر مشابه هستند می‌دهد. این الگوریتم در حوزه ارتباط رکوردها[۸] و تشخیص اسناد و موجودیت‌های تکراری[۹] مورد استفاده قرار می‌گیرد. این معیار فاصله برای رشته‌های با طول کوتاه مانند اسامی و نام‌های خاص، عملکرد بهتری دارند. پس از اعمال الگوریتم، طبق یک نرمال سازی نتیجه به صورت ۰ که مشخصه عدم شباهت و ۱ شباهت کامل نمایش داده می‌شود [۱].

معیار فاصله جارو که با d_j در فرمول زیر نمایش داده شده است برای رشته ورودی S_1 و S_2 به صورت زیر نمایش داده شده است.

$$dj = 1/3\left(\frac{m}{|s1|} + \frac{m}{|s2|} + \frac{m-t}{m}\right) \quad (۱-۲)$$

---

۱ **Insertion**

۲ **Deletion**

۳ **Substitution**

۴ **transpose**

۵ **Jaro**

۶ **Jaro-Winkler**

۷ **prefix scale**

۸ **record linkage**

۹ **duplicate detection**

که در آن m تعداد کاراکترهای همخوان و t نصف تعداد جابجایی‌ها است. دو کاراکتر از رشته‌های S_1 و S_2 به ترتیب به صورت مطابق با هم در نظر گرفته می‌شوند. اگر فاصله آنها در دو رشته بیش از مقدار زیر نباشد.

$$d = \left\lfloor \frac{max(|s1|,|s2|)}{2} \right\rfloor - 1 \quad (۲-۲)$$

هر کاراکتر از رشته S_1 با تمام کاراکترهای همسان[۱] رشته S_2 مقایسه می‌شود. برای مثال در دو کلمه 'CREATE' و 'TRACE' تنها کاراکترهای R و A و E به عنوان کاراکتر همسان در نظر گرفته می‌شوند. بنابراین مقدار $m=3$ می‌شود. با اینکه C و T در دو ورودی دیده می‌شود اما فاصله آنها بیش از ۱ است. مقدار ۱ نیز از رابطه فوق، به صورت زیر به دست می‌آید [۷].

$$\left\lfloor \frac{5}{2} \right\rfloor - 1 = 1$$

معیار فاصله جارو-وینکلر از یک درجه پیشوند با عنوان اختصاری P استفاده شده است که نسبت به جارو عملکرد بهتری دارد. این الگوریتم برای رشته‌هایی که در پیشوندشان تا فاصله L کاراکتر با هم یکسان هستند ارزش بالاتری قائل می‌شود [۸].

با دا شتن دو ر شته ورودی S_1 و S_2 فا صله جارو-وینکلر که با d_w نمایش داده می شود به صورت رابطه (۲-۳) می‌باشد.

$$d_w = d_j + (L_P (1 - d_j)) \quad (۲-۳)$$

[۱] **matching characters**

در رابطه فوق d_j فاصله جارو برای دو رشته ورودی S_1 و S_2 می‌باشد که در روابط قبلی تشریح شد.

L طول پیشوند مشترک[۱] که از ابتدای رشته آغاز می‌شود و معمولاً کمتر از ۴ کاراکتر است. P مقدار ثابت مختوم‌کننده[۲] نام دارد که مقدار وزن پیشوند مشترک را دخیل می‌کند. همچنین طبق منابع P باید بیش از ۰,۲۵ باشد. به این دلیل که اگر از این مقدار تجاوز کند فاصله از ۱ بیشتر خواهد شد (در صورتی که ما آن را ۱ و ۰ تعریف کردیم.) و همچنین مقدار استاندارد برای P برابر ۰,۱ می‌باشد.

در برخی از برنامه‌های توسعه داده شده با استفاده از جارو-وینکلر مقدار جایز $L_P\ (1 - d_j)$ تنها هنگامی در معادله دخیل می‌شود که رشته‌های مقایسه شده مقدار فاصله جارو بیشتر از یک پارامتر b_t داشته باشند. معمولاً این مقدار را ۰,۷ در نظر می‌گیرند. پارامتر b_t نیز از عبارت افزایش آستانه[۳] سرچشمه گرفته است. رابطه افزایش آستانه با خروجی جارو-وینکلر در رابطه (۴-۲) آمده است.

$$d_w = \begin{cases} d_j & if\ d_j < b_t \\ d_j + \left(L_p(1 - d_j)\right) & otherwise \end{cases} \qquad (۴-۲)$$

۴. الگوریتم n-گرام[۴]:

این الگوریتم یک زیر سری از n آیتم از رشته ورودی تشکیل می‌دهد. الگوریتم‌های مشابهت مبتنی بر n-گرام با مقایسه گرام‌های هر کاراکتر یا متن در دو رشته، فاصله را محاسبه می‌نمایند. در نهایت فاصله

---

۱ **common prefix**

۲ **sealing factor**

۳ **boost threshold**

۴ **N-Gram**

با تقسیم ساده n-گرام‌های مشابه بر تمامی n-گرام‌ها به دست می‌آید [۹]. در بخش‌های بعدی به تفصیل به کاربردهای n-گرام در محاسبه شباهت خواهیم پرداخت.

۲-۱-۱-۲- معیارهای شباهت مبتنی بر عبارات[۱]

در این بخش الگوریتم‌هایی را مورد بررسی قرار می‌دهیم که واحد اندازه‌گیری اَتمیک آنها به جای کاراکترها، کلمات و عبارات هستند. این معیارها در ابتدا رشته‌ها را به سری توکن[۲]هایی تبدیل می‌کند و سپس از معیارهای شباهت مبتنی بر عبارت برای تشخیص و محاسبه میزان شباهت استفاده می‌کنند. معیارهای مبتنی بر توکن برای رشته‌های طولانی همانند اسناد مناسب هستند. دو استراتژی به شکل رایج برای تبدیل رشته‌ها به سری‌ها استفاده می‌شوند:

۱. پیش‌پردازش و تبدیل ورودی به توکن‌ها
۲. تبدیل توکن‌ها به n-گرام

در بخش اول، رشته‌ها بر اساس کاراکترهایی خاص توکن‌بندی می‌شوند. به عنوان مثال کاراکتر فضای خالی[۳] یکی از معیارها برای جداسازی توکن‌ها به شمار می‌رود. بخش دوم از زیر رشته‌های با طول q از یک رشته برای ایجاد سری استفاده می‌کند، که زیر رشته‌ای با طول q با نام q-گرام شناخته می‌شود.

---

۱ **Term-based similarity**

۲ **Token**

۳ **White Space**

برای سادگی، هر عنصر در یک سری (کلمه یا گرام) یک توکن نامیده می‌شود و ما از رشته S برای نشان دادن سری توکن مربوطه استفاده می‌کنیم. معیارهای شناخته شده مبتنی بر توکن شامل معیار همپوشانی[۱]، جاکارد[۲]، معیار فاصله کسینوسی[۳]، و Dice هستند که در ادامه مفصلا به هر یک از این معیارها می‌پردازیم.

۱. فاصله بلوک[۴]:

این الگوریتم با نام فاصله منهتن نیز شناخته می‌شوند. در این الگوریتم فاصله بین المان‌ها که به صورت بلوک‌های شطرنجی تعریف می‌شوند بر اساس رابطه (۵-۲) محاسبه می‌شود [۱].

برای دو بردار $V = (v_1, v_2, v_3, \ \dots, v_n)$ و $W = (w_1, w_2, w_3, \ \dots, w_n)$ فاصله منتهن دو بردار برابر است با:

$$(llv - wll) = \sum_{i=1}^{n} |v_i - w_i| \quad (۲-۵)$$

۲. فاصله کسینوسی:

این معیار یک روش ارزیابی بر اساس بردار ارائه می‌کند. در این الگوریتم دو بردار ورودی در مدل فضای برداری[۵] نمایش داده می‌شود و سپس با اندازه‌گیری کسینوس زاویه‌ای ایجاد شده بین دو بردار فاصله بین این دو ورودی محاسبه می‌شود. در رابطه (۶-۲) معیار فاصله کسینوسی برای دو بردار ورودی A و B نمایش داده شده است.

$$Cosine\ Similarity = \cos(\theta) = \frac{A.B}{|A|\ |B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (۲-۶)$$

---

۱ **Overlap**

۲ **Jaccard**

۳ **Cosine Similarity**

۴ **Block Distance**

۵ **Vector Space Model**

در رابطه فوق A_i و B_i درایه‌های (توکن‌های) بردار A و B هستند. بدیهی است که هر چه میزان زاویه ایجاد شده بین دو بردار کمتر باشد کسینوس زاویه به یک نزدیک‌تر خواهد بود و دو بردار شباهت بیشتری خواهند داشت.

۳. ضریب Dice:

این معیارهر چه بیشتر باشد نشان می دهد که دو متن بیشتر به هم شباهت دارند. فاصله Dice به صورت زیر تعریف می شود.

$$Dice\ (P_1, P_2) = \frac{2 * |\ P_1 \cap P_2\ |}{|P_1| + |P_2|} \quad (۷-۲)$$

P_1 و P_2 را پروفایل می‌نامیم. این پروفایل‌ها مجموعه‌ای از بردارها هستند و یا اینکه می توان با استفاده از n-گرام‌ها متن ورودی را به پروفایل تبدیل نمود [۱۰].

۴. فاصله اقلیدسی[۱]:

فاصله اقلیدسی بر اساس ریشه اختلاف المان‌های دو بردار ورودی عمل می‌نماید و همانند فاصله منهتن از فضای برداری دو رشته ورودی استفاده می‌کند. رابطه معیار فاصله اقلیدسی به صورت رابطه (۸-۲) داده می‌شود.

$$(llv - wll_2) = (\sum_{i=1}^{n} (v_i - w_i)^2)^{\frac{1}{2}} \quad (۸-۲)$$

[۱] **Euclidean Distance**

طبیعی است که هرچقدر این مقدار بیشتر باشد نشان دهنده فاصله بیشتر دو بردار ورودی است. همچنین لازم است رابطه دیگری را معرفی کنیم که همانند رابطه اقلیدسی رفتار می‌کند. این رابطه فاصله مینکوفسکی[۱] نام دارد و برای هر q به صورت دلخواه به صورت زیر نمایش داده می‌شود [۱۱].

$$(\sum_{i=1}^{n}(|v_i - w_i|)^q)^{\frac{1}{q}} \quad (۹-۲)$$

۵. معیار شباهت جاکارد[۲]:

این معیار نیز مانند معیار Dice بر اساس تئوری مجموعه‌ها رفتار می‌کند. این فاصله به صورت تعداد المان‌های مشترک بر تعداد کل المان‌های هر دو بردار ورودی تعریف می شود [۱۲].

$$Jaccard(P_1, P_2) = \frac{|P_1 \cap P_2|}{|P_1|+|P_2|-|P_1 \cap P_2|} \quad (۱۰-۲)$$

۶. ضریب همپوشانی:

این معیار نیز مانند معیار Dice و جاکارد می‌باشد. با این تفاوت که اگر یک رشته زیر رشته‌ای از رشته دیگر باشد آنها را کاملا یکسان و متشابه در نظر می‌گیرد [۴].

$$Overlap(P_1, P_2) = \frac{|P_1 \cap P_2|}{\min(|P_1|,|P_2|)} \quad (۱۱-۲)$$

[۱] **Minkowski distance**

[۲] **Jaccard Similarity**

۲-۱-۲- معیارهای شباهت مبتنی بر پیکره[۱]

معیارهای شباهت مبتنی بر پیکره یکی از معیارهای محاسبه شباهت معنایی است که شباهت بین دو عبارت را بر اساس معنا و محتوای بدست آمده از یک پیکره بزرگ انجام می‌دهد. یک پیکره[۲] مجموعه‌ای از متون نوشته شده و یا مجموعه‌ای گفتاری است که درون تحقیقات مرتبط با زبان مورد استفاده قرار می‌گیرد. معیارهای شباهت مبتنی بر پیکره معمولا وابسته به زبان پیکره می‌باشند [۵].

۲-۱-۳- معیارهای شباهت مبتنی بر دانش

تعیین شباهت مبتنی بر دانش یکی از معیارهای تعیین شباهت است که بر پایه شناسایی درجه شباهت بین کلمات درون یک شبکه معنایی می‌باشد. WordNet معروف ترین شبکه معنایی در حوزه تعیین شباهت با استفاده از پایگاه دانش است [۱۳]. WordNet یک پایگاه بزرگ لغوی زبان انگلیسی است که اسامی، افعال، صفات و قیود در یک مجموعه مترادف شناختی[۳] گروه بندی می‌شود که اختصاراً به آن Synset گفته می‌شود. هرکدام از این مجموعه‌ها یک مفهوم مشخص را بیان می‌کنند. این مجموعه‌ها به وسیله ارتباطات لغوی و معنای مفهومی[۴] آنها بهم متصل می‌شوند [۵].

تعیین شباهت مبتنی بر دانش به طور تقریبی به دو گروه مجزای معیار شباهت معنایی[۵] و ارتباط معنایی[۶] تقسیم می‌شود. معیار شباهت معنایی بر اساس قرابت معنایی دو ورودی در شبکه واژگان و یا لغتنامه‌ها

---

۱ **Corpus based similarity**

۲ **Corpus**

۳ **Cognitive Synonym**

۴ **Conceptual Semantic**

۵ **Semantic Similarity**

۶ **Semantic Relation**

شباهت را محاسبه می‌کند. اما ارتباط معنایی[۱] الزاماً به شکل و فرم دو عبارت یا مفهوم آن‌ها وابسته نیست. به عبارت دیگر شباهت معنایی نوعی از ارتباط بین دو کلمه است که یک محدوده از ارتباطات بین مفاهیم مختلف که شامل ارتباطات معنایی دیگری از قبیل "نوعی از"[۲]، "نمونه‌ای از"[۳]، "بخشی از"[۴] و "متضاد با"[۵] را شامل می‌شود [۵].

۲-۲- بررسی روش‌های لغوی محاسبه شباهت بین دو رشته

در قسمت پیشین بر روش‌های مختلف محاسبه مشابهت میان اسناد مروری کلی داشتیم. در این قسمت ما روی مشابهت بین رشته‌ها در یک محیط تک‌زبانه و روش‌های مستقل از زبان تمرکز خواهیم داشت. لازم به ذکر است که برخی از این روش‌ها در تشخیص سرقت علمی-ادبی کاربرد دارند. از این حیث مقالات مرتبط در زمینه تشخیص سرقت علمی ادبی نیز مورد بررسی قرار گرفته اند.

۲-۲-۱- روش مبتنی بر فشرده‌سازی دو رشته

روش‌های مبتنی بر تکنیک‌های فشرده سازی فایل‌ها، در حقیقت به صورت مستقیم از الگوریتم‌های فشرده سازی کمک می‌گیرند. در این روش دو سند به یکدیگر الحاق شده و سپس فشرده‌سازی می‌شود. همچنین هر دو سند نیز به صورت جداگانه فشرده می‌شوند و در نهایت حجم فایل فشرده‌سازی شده با مجموع حجم فایل‌های فشرده‌سازی به به صورت جداگانه مقایسه می‌شود. اگر دو سند متفاوت باشند اندازه فایل حاصل از دو حالت برابر خواهد بود. اما اگر این فایل‌ها حاوی قسمت‌های تکراری باشند اندازه حالت اول کوچکتر

---

۱ **Semantic Relation**

۲ **Is-A-Kind-Of**

۳ **Is-A-Specific-Example-Of**

۴ **Is-A-Part-Of**

۵ **Is-The-Opposite-Of**

می‌شود. از مزیت‌های این روش سادگی و سهولت پیاده‌سازی است. اما این روش نمی‌تواند حالات پیچیده و جابجایی‌های کلمات و جملات را تشخیص دهد [۱۴].

۲-۲-۲- مدل فضای بردار[۱] (VSM)

مدل فضای بردار روشی بر اساس محاسبه فرکانس کلمه و تشکیل بردار وزنی TF-IDF[۲] بر اساس آن می‌باشد. شباهت بین بردار وزن از دو سند با استفاده از یکی از معیارهای شباهت بردار انجام می‌گیرد. اطلاعات به شکل بردار در فضای چند بعدی نشان داده می‌شوند. هر بعد مربوط به یکی از ویژگی‌های اطلاعاتی مانند هر کلمه در سند می‌باشد. یک تابع فاصله روی بردارها برای محاسبه انطباق و رتبه‌بندی اطلاعات اعمال می‌گردد که می‌توان از هر یک از روابط گفته شده در بخش شباهت مبتنی بر عبارت استفاده کرد. این مدل گاهی در کنار یک معیار شباهت مانند شباهت کسینوسی مورد استفاده قرار می‌گیرد [۱۵].

بازیابی اطلاعات مبتنی بر مدل فضای بردار یک پیاده سازی ریاضی مناسب برای پردازش منابع اطلاعاتی بزرگ است و امکان تطبیق جزئی و رتبه‌بندی خروجی را فراهم می‌کند. با این حال، این روش فاقد ایجاد روابط معنایی است و هیچ ابزاری برای مدیریت معنایی یا نحوی اطلاعات ندارد. امروزه، بسیاری از روش‌هایی که برای نمایش اسناد استفاده می‌شوند بر مدل کیسه کلمات[۳] (BOW) تکیه می‌کنند، که به طور معمول به عنوان مدل فضای برداری شناخته می‌شود [۱۶]. اسناد به صورت یک بردار خطی از وقوع کلمات در پیکره متنی سند نشان داده می‌شوند. از معایب دیگر روش فوق می‌توان به این موارد اشاره کرد که بسیاری از روابط معنایی میان مفاهیم و اطلاعات مهم، هنگام استفاده از مدل فضای برداری از دست می‌روند. و

---

۱ **Vector Space Model**

۲ **Term frequency – inverse document frequency**

۳ **Bag of word**

همچنین اگر سند طولانی باشد، نمایش آن به صورت مدل برداری با توجه به ابعاد زیاد بسیار دشوار می‌باشد و حالت انفجار فضای برداری رخ خواهد داد [۱۷].

۲-۲-۳- محاسبه شباهت بین پاسخ‌ها در سیستم‌های پرسش و پاسخ

این روش در پژوهش مربوط به محاسبه شباهت بین پاسخ‌ها در پورتال‌های پاسخ آنلاین [۴] منتشر شده است. این پژوهش روشی مبتنی بر شکل لغوی جملات و عبارات تشکیل‌دهنده یک جمله ارائه می‌دهد. در سیستم پرسش و پاسخ، از الگوریتم‌های محاسبه شباهت بین عبارات برای شباهت بین دو سوال یا بین دو پاسخ استفاده می‌شود. با استفاده از معیارهای متفاوت شباهت، به ویژه معیارهای شباهت معنایی و معیارهای تشابه آماری، می‌توان به این هدف دست یافت. بنابر حجم زیاد تعداد پرسش و پاسخ‌ها پیاده‌سازی سنجش‌های شباهت معنایی بسیار پیچیده و زمانبر می‌باشد و استفاده از الگوریتم‌های شباهت آماری مقرون به صرفه‌تر است. تمرکز اصلی در اینجا این است که تشابه پاسخ کاربر را با پاسخ ذخیره شده در پایگاه داده پرسش و پاسخ بهبود بخشیده.

بحث‌هایی در مورد سنجش شباهت جاکارد انجام شده است که در آن شباهت بر اساس تعداد کلمات مشترک بین دو جمله محاسبه می‌شود. این، شباهت بین دو مجموعه کلمات را مقایسه می‌کند. اگر S_a و S_b دو جمله باشند، پس برای مقایسه، نیاز است که در ابتدا مجموعه کلمات دو جمله آماده شوند. ساختار مجموعه کلمات شامل نشانه‌گذاری[۱] جملات است که در آن جملات به کلمات تقسیم می‌شوند. فرض کنید که S_a و S_b دو جمله به شرح زیر هستند.

S_a = سبزیجات سبز را دوست ندارم. S_b = من آنها را دوست ندارم.

[۱] **tokenization**

پس از نشانه‌گذاری، S_a و S_b به شرح زیر می‌شود:

S_a = ["من"، "سبزیجات"، سبز"، "را" ،" دوست"، "ندارم"].

S_b = ["من"، "آنها"، "را"، "دوست"، ندارم"].

از موارد فوق اطلاعات زیر به دست خواهد آمد.

جدول ۲-۱ مثال واقعی سیستم پرسش و پاسخ در ارزیابی معیار جاکارد

پارامتر	مقدار
تعداد کل کلمات موجود در S_a	۶
تعداد کل کلمات موجود در S_b	۵
تعداد کل کلمات موجود مشترک در S_a و S_b	۴
تعداد کل کلمات مختلف موجود در S_a و S_b	۷
شباهت جاکارد	۰,۵۷

معیار تشابه جاکارد برای محاسبه شباهت، معیار بسیار کارآمدی است، اما به مسئله دقت پاسخ‌ها در سیستم‌های پرسش و پاسخ آنلاین نمی‌پردازد. برای بهبود این روش از معیار همپوشانی می‌توان استفاده کرد. این معیار عبارات مشترک بین دو ورودی را مدنظر قرار می‌دهد. در نهایت معیار شباهت به صورت زیر خواهد بود:

$$Overlap\ Similarity = \frac{W_{S_a} \cap W_{S_b}}{\min(W_{S_a}, W_{S_b})} \quad (۱۲-۲)$$

به طور خاص، در پورتال پرسش‌پاسخ چندین سوال وجود دارد که برای آنها پاسخ می‌تواند در یک کلمه، یک خط یا در یک پارگراف داده شود. به عنوان مثال برای سوال "پایتخت هند چیست؟" می‌توان آن را به سه روش پاسخ داد: ۱. پایتخت هند دهلی است. ۲. دهلی. ۳. دهلی پایتخت هند است.

همه این پاسخ‌ها با توجه به سوال داده شده درست هستند. اگر شباهت با روش ساده جاکارد محاسبه شود، شباهت قطب‌هایی جدا از هم هستند و بنابراین در این روش، پیشرفتی در جهت برطرف کردن مسئله فوق‌الذکر صورت گرفته است. در این فرآیند، اگر تمام کلمات (توکن‌ها) پاسخ کاربر در پاسخ ذخیره‌شده در پایگاه داده پرسش و پاسخ موجود باشند، شباهت یک در نظر گرفته می‌شود که همگی آنها پاسخ درست تلقی می‌شوند [۴]. مجموعه داده برای بررسی این روش، مجموعه‌ای از پرسش‌ها درنظر گرفته شده است که پاسخ صحیح برای هر سوال در پایگاه داده پرسش و پاسخ ذخیره می‌شود. در مجموعه داده، طول پاسخ‌ها بین ۱۰ تا ۲۰ کلمه است. در این روش پاسخ‌هایی با حداقل و حداکثر اندازه طول انتخاب می‌شوند، زیرا تمرکز بر روی بررسی پاسخ ذهنی است.

۲-۲-۴- روش‌های مبتنی بر اثر انگشت[۱]

روش‌های مبتنی بر الگوریتم اثر انگشت [۱۸] و هش [۱۹] از روش‌های مبتنی بر محتوا به شمار می‌روند. روش‌های مبتنی بر محتوا به مقایسه صریح و روشن محتویات سند در یک نمایش خاص می‌پردازد. اثر انگشت و مدل‌های مبتنی بر هش جزو مدل‌های بازیابی اکتشافی هستند که در آنها یک مجموعه از اعداد صحیح از طریق هش کردن زیر مجموعه‌ای از سند، برای نشان دادن کلید محتوای آن ایجاد می‌شود. فرآیند ایجاد اثر انگشت، انگشت‌نگاری[۲] نامیده می‌شود. اثر انگشت سند می‌تواند سند را به صورت منحصر بفرد شناسایی کند، همانند اثر انگشت در انسان [۲۰]. n-گرام به عنوان الگوریتم پایه برای بسیاری از روش‌های اثر انگشت استفاده می‌شود، زیرا فرآیند انگشت‌نگاری سند را به گرام‌هایی با طول مشخص n تقسیم می‌کند. مدل مبتنی بر هش از یک تابع هش[۳] برای تبدیل اثر انگشت به مقدار هش استفاده می‌کند که می‌توان آن

---

۱ **Finger print**

۲ **Fingerprinting**

۳ **Hashing Function**

را ذخیره کرد و با هش سایر اسناد مقایسه کرد. تابع هش، یک تابع یک به یک می‌باشد که رشته‌ها را به یک فضای عددی ۳۲ رقمی در مبنای ۱۶ تبدیل می‌کند. هر ورودی این تابع به یک عدد منحصر به فرد[۱] تبدیل می‌کند. در نهایت با تبدیل تمامی توکن‌ها به هش، لیستی از اثر انگشت مربوط به هر سند، بردار آن سند را می‌سازد. متریک‌های شباهت بردار را می‌توان برای بازیابی اسنادی که اثر انگشت آنها تولید شده استفاده کرد. گرام‌ها در واقع جزء ویژگی‌های واژگانی محسوب می‌شود که این ویژگی‌های واژگانی می‌توانند در سطح کاراکتر یا کلمه عمل کنند. به این صورت که این ویژگی‌ها را می‌توان به دو شکل مختلف n-گرام مبتنی بر کاراکتر[۲] (CNG) که دنباله‌ای از کاراکترها است، و n-گرام مبتنی بر کلمات[۳] (WNG) که دنباله‌ای از مجموعه کلمات با نادیده گرفتن جملات و مرزهای ساختاری است در نظر گرفت. WNG ساده ممکن است به صورت گرام دوتایی یا گرام سه تایی یا بزرگتر ساخته شود. در بازیابی متن و تحقیقات محاسبه شباهت CNG و WNG عموما به نام اثر انگشت یا شینگل[۴] نامیده می‌شوند [۲۱]. در CNG طبق تحقیق‌های پیشین از ۸ و یا ۱۶ گرام استفاده می‌شود. در این روش مرز بین تمامی عبارات و جملات از بین می‌رود و متن به یک سری پشت‌سرهم از کاراکترها تبدیل می‌شود. پس از آن از رشته حاصله n-گرام کاراکترها استخراج می‌شوند. [۲۲] WNG ساده ممکن است به صورت گرام‌های ۲ تا ۵ تایی تشکیل شود [۲۳].

در تکنیک اثر انگشت می‌توان صورت مسئله را اینگونه تعریف کرد: همپوشانی دو سند که تعداد مشخصی اثر انگشت مشترک داشته باشند به گونه‌ای که اثر انگشت‌ها ویژگی‌های اسناد را به طور کامل در خود جای داده باشند [۲۴].

---

۱ **Unique**

۲ **Character-based n-gram**

۳ **Word-based n-gram**

۴ **Shingle**

مزیت اصلی روش انگشت‌نگاری سریع بودن آن است و می‌تواند به طور موثر در مجموعه‌های بزرگ استفاده شود. همچنین این روش معایبی نیز به همراه دارد. یکی از معایب این روش در نظر نگرفتن اطلاعات معنایی جملات و عبارات می‌باشد. با این وجود تطبیق اثرانگشت از نظر زمان و فضای حافظه هزینه‌بر است و از این روش در تطبیق‌های دقیق استفاده می‌شود [۲۲].

تکنیک‌های انتخاب اثر انگشت با توجه به طرح‌های مختلف انجام می‌گیرد. در جدول (۲-۲) این روش‌ها در کنار یکدیگر مقایسه شده اند.

جدول ۲-۲ مقایسه روش‌های انتخاب شینگل در اثر انگشت [۲۵]

نام	روش کار	مزایا و معایب
i امین هش	همه iامین هش‌ها از یک سند انتخاب می‌شوند.	پیاده‌سازی آن آسان اما در صورت درج، حذف یا تغییر ساختار متن قوی نیست.
هش مد p	در آن p یک عدد صحیح است، مکان‌های هش که مد p صفر است، انتخاب می‌شوند.	پیاده‌سازی آسان است. محتوای مشابه اگر رشته هش آن در مکان مد p باشد و انتخاب شود تشخیص داده می‌شود.
پنجره کشویی	تعریف یک پنجره به صورت متوالی و به اندازه W که شامل هش‌ها است. در هر پنجره تعداد مشخصی از کوچکترین هش‌ها انتخاب می‌شود.	انتخاب به صورت نرمال انجام می‌گیرد و از تمامی هش کاندید انتخاب می‌شود. در تشخیص یک جمله (سری متوالی از کلمات) به سرقت رفته عملکرد خوبی دارد.

۱-۴-۲-۲- روش Winnowing

یکی از معروف‌ترین الگوریتم‌های انگشت‌نگاری محلی توسعه یافته روش winnowing است. غالب سیستم‌های تحلیل و بررسی سرقت علمی-ادبی بر پایه این الگوریتم عمل می‌کنند. این الگوریتم از مراحل زیر تشکیل شده است:

۱. حذف تمامی اطلاعات نامربوط از متن، که حذف این اطلاعات در دقت و کارایی الگوریتم تاثیر بسزایی دارد.
۲. ایجاد k-گرام از اطلاعات بدست آمده از مرحله اول. در اینجا k تا از حروف را در نظر گرفته و گرام بعدی با یک حرکت به سمت را ست در ر شته بد ست می‌آید. این کار تا ر سیدن به پایان ر شته بدست آمده در مرحله اول ادامه داده می‌شود. کلمات تولید شده در اصطلاح گرام‌هایی هستند که مبنای ارزیابی و مقایسه اسناد می‌باشند.
۳. k-گرام‌های تولید شده توسط یک الگوریتم مناسب به هش تبدیل شده. دلیل کار با هش به این علت است که به جای کار با رشته بتوان عملیات را با اعداد انجام داد.
۴. از هش‌های بدست آمده نمونه‌برداری صورت می‌گیرد. این نمونه‌ها ذخیره و آنها اثر انگشت تمام متن می‌باشند [۲۵].

این الگوریتم اصطلاحا تکه[۱] یا پنجره کشویی[۲] نیز نامیده می‌شود. تکه‌بندی[۳] جزء ویژگی‌های لغوی محسوب می‌شود و شامل برش متن به قطعات کوچکتر است که برای توصیف قطعه متون بزرگ‌تر مانند عبارات یا

---

[۱] **Chunk**

[۲] **Sliding windows**

[۳] **Chunking**

جملات به کار می‌رود [۲۶]. در تکه‌بندی مبتنی بر جمله، سند به تکه‌هایی بر اساس پارامتر n تقسیم می‌شود و هر دنباله از n جمله در یک تکه گروه‌بندی می‌شوند.

برای بهبود عملکرد winnowing طول پنجره‌ها به صورت $t-k+1$ در نظر گرفته می‌شود. در این حالت t وk آستانه تضمین و آستانه نویز هستند. برای پیدا کردن مطابقت بین دو سند دو ویژگی باید تضمین گردد:

۱. مطابقت تشخیص داده خواهد شد اگر یک زیر رشته مشابه حداقل به طول آستانه تضمین t وجود داشته باشد.
۲. هر تطابق کوچکتر از آستانه نویز k تشخیص داده نخواهد شد [۲۵].

تحقیقات نشان می‌دهد روش تطبیق اثر انگشت با توجه به واحد مقایسه گرام متفاوت خواهد بود. در نتیجه می‌تواند در سه دسته طبقه‌بندی شود:

۱. در اثر انگشت مبتنی بر کاراکتر از گرام‌های حدود ۱۶ تایی استفاده می‌شود.
۲. اثر انگشت مبنی بر عبارت، برای اندازه‌گیری شباهت دو سند از اثر انگشت مبتنی بر عبارت استفاده می‌شود. تشخیص شباهت با استفاده از انگشت‌نگاری مبتنی بر عبارت بهتر و سریع‌تر از انگشت‌نگاری مبتنی بر کاراکتر است، زیرا از کلمات به جای حروف استفاده می‌کند. با این حال، این روش برای اسناد با اندازه بزرگ زمان و فضای زیادی مصرف می‌کند [۲۷].
۳. اثر انگشت مبتنی بر جمله، اثر انگشت را برای هر جمله با سه حداقل تکرار ۴-گرام‌ها در نظر می‌گیرند. مزایای استفاده از این روش پردازش سریع و مصرف کمتر حافظه است. با این حال، در مورد جایگزینی کلمات و بازسازی جملات در این روش، شکست خواهد خورد، اما در محاسبه سرقت ناشیانه مبتنی بر جملات موثر است [۲۷].

۲-۲-۴-۲- روش SimHash

بیشتر الگوریتم‌های مبتنی بر تکنیک اثر انگشت از تمام متون اثر انگشت‌های بسیار زیادی تهیه می‌کنند. برای مواردی کاربردی همچون خزنده‌های وب[۱] و سایر الگوریتم‌های بلادرنگ که زمان و حافظه نقش بسیار مهمی در آن دارد، لازم است محاسبه شباهت توسط الگوریتم نسبتا سریع اجرا شود. الگوریتم SimHash به علت عملکرد و ساختار الگوریتمیک برای حجیم‌داده‌ها کارایی نسبتا مناسبی دارد [۲۸].روش SimHash نسبت به سایر الگوریتم‌های مبتنی بر هش و اثر انگشت، حافظه کمتری مصرف می‌کند. همچنین اینکه سرعت این الگوریتم نسبت سایر الگوریتم‌ها مانند winnowing سریع‌تر است.

روش محاسبه شباهت SimHash در سال ۲۰۰۲ توسط آقای چاریکار[۲] ارائه شد و در سال ۲۰۰۷ توسط چند موتور جستجوی معروف از جمله گوگل و آلتاویستا[۳] به کار گرفته شد. مزیت این روش مصرف حافظه کم و سرعت بالای آن است.[۲۸] در این روش می‌توان چندین اثر انگشت را به یک اثر انگشت نگاشت داد. مزیت SimHash نسبت به سایر الگوریتم‌های هش این است که خروجی SimHash برای دو متن مشابه، خروجی مشابهی را دارد، در حالی که خروجی هش برای دو متن بسیار مشابه و با اختلاف حتی یک کاراکتر بسیار متفاوت خواهد بود [۲۹].

در این مرحله در ابتدا بایستی یک توالی از n-گرام کلمات ساخته شود. در زبان فارسی معمولا از ۳ و ۴ برای تولید n-گرام‌ها استفاده می‌شود. بنابراین اگر متن ورودی دارای N توکن باشد، پس از استخراج n-گرام‌ها شامل N-n+1 گرام با اندازه هر گرام n خواهد بود. حال بایستی یک تابع هش روی این ویژگی‌های

---

۱ **Web Crawlers**

۲ **MS Charikar**

۳ **Alta Vista**

تولید شده اعمال گردد. سه روش روش هش‌نگار مرسوم رابین[۱]، MD5 و SH2 برای تولید هش‌ها وجود دارند.

پس از استخراج ویژگی‌ها و دریافت مجموعه‌ای از اعداد ۳۲ بیتی حاصل از اعمال تابع هش روی ویژگی‌ها، اعداد به مبنای دودویی نگاشت داده می‌شوند. حال مجموعه ۱۲۸ بیتی از اعداد در اختیار است. در ادامه کار یک بردار ۱۲۸ بعدی با مقدار اولیه ۰ ایجاد می‌شود. این بردار v در نظر گرفته می‌شود. برای هر ویژگی خواهیم داشت:

- اگر بیت iام یک باشد، به $v[i]$ یک واحد اضافه می‌کنیم.
- اگر بیت iام صفر باشد، از $v[i]$ یک واحد کم می‌کنیم.

در نهایت برای کل مجموعه ویژگی‌ها تنها با یک بردار ۱۲۸ بعدی مواجه هستیم. حال از روی این بردار خروجی SimHash را می‌توان تعیین کرد، به این صورت که اگر $v[i]$ بزرگ‌تر از صفر باشد، خروجی $SimHash[i]$ یک می‌شود، در غیر این صورت صفر خواهد شد. با استفاده از این تکنیک همه متن به یک بردار ۱۲۸ بیتی تبدیل خواهد شد. برای تشخیص میزان شباهت دو متن فاصله همینگ بین آنها محاسبه خواهد شد. هرقدر این مقدار کمتر باشد دو متن بیشتر به یکدیگر شباهت دارند [۳۰].

۲-۳- تکنیک‌های انطباق رشته‌های مشابه[۲]

انطباق رشته‌های مشابه، یک عملیات مهم در یکپارچه‌سازی داده و پاکسازی پایگاه‌های داده است که در آن جفت رشته‌های مشابه را از دو مجموعه رشته می‌یابد. با داشتن دو مجموعه رشته، مثل نام محصولات و

[۱] **Rabin Hash Function**

[۲] **String similarity Joins**

فیلم‌ها، الگوریتم انطباق رشته‌های مشابه باید تمام جفت رشته‌های مشابه را از این دو مجموعه بیابد. شباهت بین دو رشته را می‌توان با معیارهای شباهت تشخیص داد. انطباق رشته‌های مشابه نقش مهمی در بسیاری از موارد در دنیای واقعی مثل پاکسازی و یکپارچه‌سازی داده، و تشخیص موارد تکراری ایفا می‌کند [۳۱].

یک تکنیک رایج از چارچوب فیلتر-اعتبارسنجی استفاده می‌کند، که شامل دو مرحله است:

۱. **مرحله فیلتر:** استفاده از الگوریتم‌های بهینه فیلتر برای هرس کردن تعداد زیادی از جفت‌های غیرمشابه و ایجاد یک سری جفت کاندیدا.

۲. **مرحله اعتبارسنجی:** اعتبارسنجی هر جفت کاندیدا با محاسبه شباهت واقعی و استخراج نتایج نهایی.

الگوریتم‌های فیلتر در مرحله اول نقش مهمی در این چارچوب دارند. بیشتر الگوریتم‌های فیلتر موجود، از تکنیکی مبنی بر امضا استفاده می‌کنند، که امضاهایی را برای هر رشته به طریقی ایجاد می‌کند که اگر دو رشته مشابه هم باشند، امضای آنها باید با هم اشتراک داشته باشد. بنابراین تکنیک مبنی بر امضا می‌تواند جفت رشته‌هایی را هرس کند که امضای مشترکی ندارند.

بسیاری از تکنیک‌های فیلتر کردن از جمله فیلتر تعداد، فیلتر طول، فیلتر موقعیت، فیلتر پیشوند و فیلتر محتوا در این زمینه ارائه شده‌اند. فیلتر پیشوند بهینه‌ترین تکنیک فیلتر کردن است. از آنجایی که فیلتر پیشوند بهینه‌ترین تکنیک فیلتر کردن است بسیاری از الگوریتم‌ها برای بهینه‌سازی فیلتر پیشوند برای معیارهای مختلف شباهت ارائه شده‌اند که مهم‌ترین آنها شامل Probe-Count، AllPair، PPJoin، EDJoin، QChunk، VChunk، AdaptJoin است [۳۲].

هدف مرحله فیلتر استفاده از الگوریتم‌های بهینه‌ی فیلتر برای هرس کردن جفت‌های نامشابه است. از آنجایی که الگوریتم فیلتر خود دارای سربار است، افزایش قدرت فیلتر به ازای افزایش هزینه در کارایی الگوریتم

خواهد بود. به عبارت دیگر، الگوریتم فیلتر باید سبک باشد در حالی که باید بتواند تعداد زیادی از جفت‌های غیرمشابه را هرس کند. علاوه بر این برای تشخیص شباهت دقیق، الگوریتم‌های فیلتر باید نقطه ضعفی نداشته باشند. برای رسیدن به این ملزومات، یک تکنیک رایج در مرحله‌ی فیلتر تکنیک مبتنی بر امضا است، که یک سری امضا را برای هر رشته به طریقی ایجاد می‌کند که اگر دو رشته مشابه هم باشند باید حداقل یک امضای مشترک داشته باشند و همچنین جفت‌های غیرمشابهی که امضای مشترکی ندارند را هرس می‌کند.

هدف مرحله اعتبارسنجی بررسی اعتبار هر جفت کاندیدا با محاسبه شباهت واقعی آنها است. این معیارهانیز به دو دسته مجزای مبتنی بر توکن و مبتنی بر کاراکتر تقسیم‌بندی می‌شوند.

۱–۳–۲- دسته‌بندی روش‌های بررسی انطباق رشته

به طور کلی بررسی انطباق رشته‌ها به دو روش مبتنی بر نمایه معکوس و فیلترینگ پیشوندی و پسوندی صورت می‌پذیرد. این روش‌ها جزو روش‌های دقیق محاسبه شباهت به حساب می‌آیند. نام روش دقیق بر روی این نوع معیارها به این دلیل بنا نهاده شده است که این روش‌ها مقدار دقیق و واقعی شباهت بین دو متن را محاسبه و اندازه‌گیری می‌کنند و بر مبنای آن عمل می‌کنند. این روش‌ها به دلیل ایکه مقدار شباهت بین دو متن را به طور دقیق محاسبه می‌کنند نسبتاً کند هستند و روش‌های مطرح در این زمینه غلب برای مقابله با این کندی با استفاده از نمایه‌های معکوس و روش فیلترینگ مناسب سعی در بهبود کارایی این روش‌ها دارند. روش‌های مبتنی بر فیلترینگ پیشوندی و پسوندی در واقع توسعه روش‌های مبتنی بر نمایه‌های معکوس هستند و با استفاده از اصول فیلترینگ پیشوندی و پسوندی سعی در کاهش اندازه نمایه معکوس و بهبود کارایی دارند [۳۲].

نمایه معکوس اصطلاحی است که در بازیابی اطلاعات مطرح است. این نمایه هر کلمه را به لیستی از رکوردها یا متن‌هایی که کلمه مورد نظر در آنها آمده است می‌نگارد. از مهم‌ترین الگوریتم‌های نمایه معکوس می‌توان به Probe-Count، Pair-Count و Probe-Stop-Words اشاره کرد. روش‌های فیلترینگ پیشوندی و پسوندی نیز بخشی از روش نمایه معکوس به حساب می‌روند، اما به دلیل رویکرد خاص این روش‌ها در کاهش حجم نمایه معکوس و روش انجام آن، این روش‌ها به طور جداگانه بررسی می‌شوند.

۲-۴- روش‌های مبتنی بر سیستم‌های استنتاج فازی

تاکنون روش‌ها و سنجه‌هایی که در محاسبه مشابهت و دستبرد ادبی دو سند بررسی شد به صورت قطعی میزان شباهت بین دو متن را محسابه می‌کردند. آهنگربهان و منتظر [۳۳] مدلی در سنجش شباهت لغوی محتوای منابع وب فارسی در شرایط عدم قطعیت ارائه کرده‌اند. با توجه به کیفی بودن عوامل اثرگذار به سنجش شباهت بین دو متن، روش جدید با استفاده از نظریه مجموعه فازی ارائه شده است. در این روش ابتدا واژه‌های هر سند به دو دسته عمومی و تخصصی بخش‌بندی می‌شوند و سپس برای هر دسته سنجه‌ها و ویژگی‌های متناسب با آن استخراج می‌شود و در نهایت با استفاده از سیستم استنتاج فازی، ارزیابی قرار می‌گیرد. در این بحث روشی برای برطرف‌سازی نقص اطلاعاتی درباره سنجش شباهت در محتوای تخصصی و علمی منابع ولی با استفاده از بیان فازی ارائه گشته است که با دقت و اطمینان بالایی شباهت لغوی دو محتوا را محاسبه می‌کند.

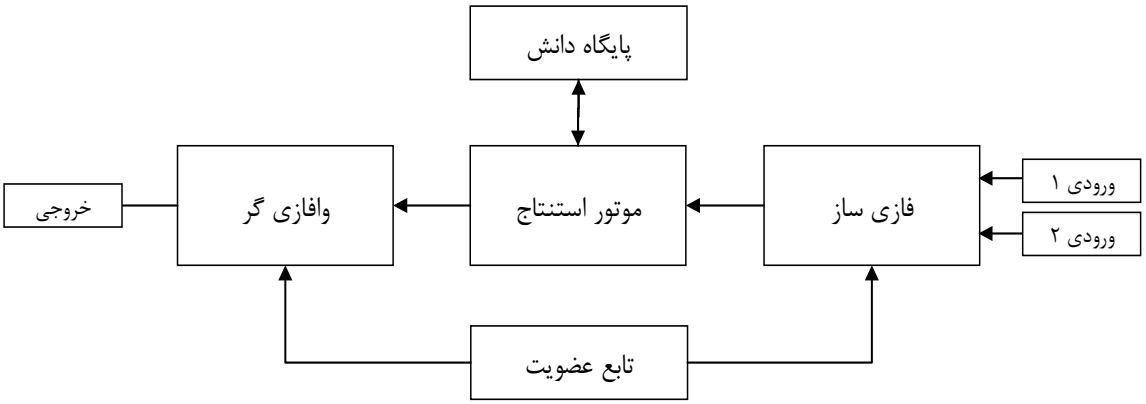
از آنجا که معمولا کاربران از تغییر ترتیب واژه‌ها و یا جابه‌جاکردن بخش‌های مختلف متن برای بازنویسی متن (به خصوص در متون تخصصی) استفاده می‌کنند [۳۴]، تشخیص این نوع محتوای مشابه و دستبرد دشوار خواهد بود و نیاز است برای به دست آوردن کارایی بهتر بیان دقیق‌تری از مسئله ارائه شود و عوامل تاثیرگذار بر سنجش و تشخیص شباهت بین دو متن شناسایی گردد.

بر اساس رویکردهای پیشین که مورد بررسی قرار گرفت استنتاج می‌شود که وجود این رویکردها نشان‌دهنده آن است که شباهت مفهمومی و ذهنی، مبهم و نامعلوم است و نیاز است بیان شفاف‌تر از شباهت ارائه گردد. از طرفی تنها یک سنجه شباهت‌سنجی برای سنجش در نظر گرفته شده است که چندان سنجش کاملی به دست نمی‌دهد. از این رو می‌توان با استفاده از چندین معیار و ویژگی و یک موتور استنتاج فازی نتایجی شبیه به نتایج خبرگان دریافت نمود. در این بخش مروری کوتاه بر بر روی سیستم استنتاج فازی خواهیم داشت. پس از آن به ذکر جزئیات سیستم استنتاج فازی محاسبه شباهت خواهیم پرداخت.

۲-۴-۱- نظریه مجموعه فازی و سیستم‌های استنتاج فازی

عدم قطعیت یکی از مواردی است که برای انسان‌ها در شناخت محیط به وجود می‌آید. انسان‌ها معمولا در تحلیل بیان کمی دچار مشکل بوده و از سوی دیگر در قضاوت کیفی به صورت کارا موفق عمل می‌کنند. پروفسور لطفی‌زاده برای حل این مشکل نظریه مجموعه فازی را ارائه داده‌اند. این نظریه، چارچوب مناسبی را برای محاسبه داده‌ها و اطلاعات غیرقطعی و مبهم ارائه کرده است. این نظریه می‌تواند روابط موضوعی و عدم قطعیت را به صورت ریاضی بیان کند. سیستم استنتاج فازی نیز بخش تصمیم‌گیرنده و مغز سیستم فازی است که فرآیند نگاشت ورودی به خروجی را با استفاده از نظریه فازی انجام می‌دهد [۳۵].

شکل (۱) نمونه‌ای از سیستم استنتاج فازی است که شامل ۵ بخش فازی‌ساز، تابع عضویت، موتور استنتاج، پایگاه دانش و فازی‌گر است را نشان می‌دهد. قلب هر سیستم فازی پایگاه دانش است که از ترکیب دانش خبرگان حوزه مورد بحث و یا به صورت بدون نظارت توسط سایر الگوریتم‌های هوش مصنوعی استخراج می‌شود. در این چارچوب استنتاج از طریق مجموعه‌ای از قواعد اگر-آنگاه انجام می‌شود که هر یک از این قواعد به کمک مجووعه‌های فازی تعریف می‌شود [۳۶].

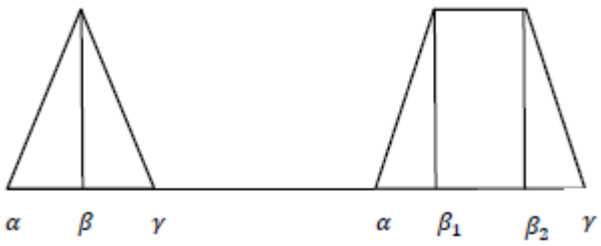


شکل ۲-۱ معماری سیستم استنتاج فازی

به صورت خلاصه هر یک از اجزای فوق را تشریح می‌نماییم:

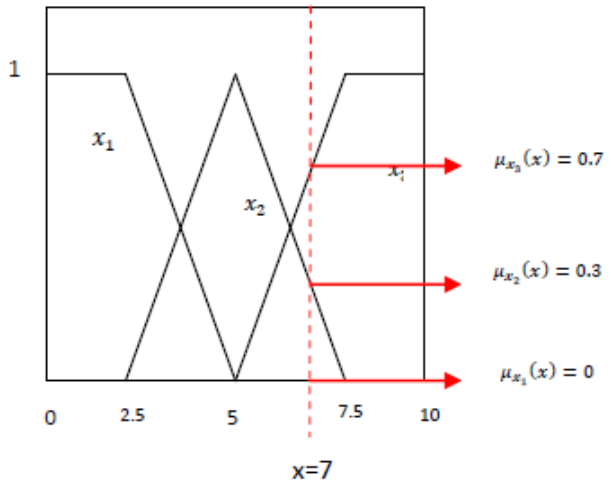
۲-۴-۱-۱- فازی ساز

وظیفه اصلی فازی‌ساز این است که در ورودی مقدار عددی متغیرها را به یک مجموعه فازی تبدیل نماید. در این مرحله برای هر متغیر ورودی، توابع عضویت در نظر می‌گیریم تا ورودی‌های قطعی تبدیل به فازی شوند و در سیستم استنتاج فازی قرار بگیرند. توابع عضویت انواع مختلفی دارند، مانند مثلثی، ذوزنقه‌ای، گوسی و غیره. تابع مثلثی و ذوزنقه‌ای که معروف‌ترین این نوع است و ما در بخش بعدی از این توابع استفاده خواهیم کرد به ترتیب به صورت $(\alpha, \beta, \Upsilon)$ و $(\alpha, \beta_1, \beta_2, \Upsilon)$ که مقداریر هر یک از این پارامترها در شکل زیر نشان داده شده‌است.



شکل ۲-۲ ساختار توابع عضویت مثلثی و ذوزنقه‌ای در سیستم‌های استنتاج فازی

عملیات فازی‌سازی بر روی توابع فوق صورت می‌پذیرد. در نهایت پس از فازی‌سازی متغیر فرضی x، بر اساس نقطه برخورد محور x با هر یک از توابع فوق تعداد یک و یا بیشتر تابع عضویت x_1، x_2 و x_n برای آن در نظر گرفته خواهد شد. به عنوان مثال با در نظر گرفتن شکل زیر به عنوان یک سری مجموعه فازی اگر مقدار x برابر ۷ باشد می‌توان گفت که متغیر x با درجه عضویت ۰ متعلق به مجموعه x_1، با درجه عضویت ۰,۳ متعلق به x_2 و با درجه عضویت ۰,۷ متعلق به x_3 است.



شکل ۲-۳ مثال درجه عضویت با وجود سه مجموعه فازی متفاوت

که به صورت ریاضی به این‌گونه بیان می‌گردد:

$$\mu x_1(x) = 0 \qquad \mu x_2(x) = 0.3 \qquad \mu x_n(x) = 0.7 \qquad (۱۴-۲)$$

۲-۴-۱-۲- پایگاه قواعد

پایگاه قواعد به مجموعه "اگر -آنگاه" فازی گفته می شود که قلب سیستم استنتاج فازی را تشکیل می‌دهد. دو روش عمده برای تعیین قواعد فازی وجود دارد:

۱. **ا ستفاده از دانش خبره:** چندین خبره بر ا ساس مجموعه ویژگی‌ها و ا ستنتاج‌های دنیای واقعی قواعدی را وضـع می‌کنند که سـیسـتم فازی از این مجموعه قواعد خروجی مورد نظر را اسـتنتاج خواهد کرد.

۲. **استفاده از آموزش‌های خود سازمانده:** مانند الگوریتم های مبتنی بر شبکه‌های عصبی است.

در این بحث از روش اول برای تعیین قواعد فازی خواهیم نمود. یک قانون اگر-آنگاه به صورت "اگر X برابر A باشد، آنگاه Y برابر B است"تعریف می شود که X و Y متغیرهای ورودی و خروجی و A و B مقادیر زبانی (توابع عضویت) نوشته شده برای این متغیرها است. قسمت "اگر X برابر A باشد"، قسمت "مقدم یا فرض" و قسمت "آنگاه Y برابر B است" را قسمت "تالی یا برآیند" گویند.

۲-۴-۱-۳- موتور استنتاج فازی

عملکرد موتور استنتاج فازی شبیه فرآیند استدلال آدمی است، به طوری که با اعمال آن بر روی ورودی‌ها و قواعد، خروجی مشخص می‌شود و این همان کاری است که انسان در بسیاری از قضاوت‌های خود به کار می‌گیرد. به این صورت که با اعمال ورودی‌ها درجه عضویت هر یک از توابع استخراج شده. به طور کلی عملکرد موتور استنتاج فازی را می‌توان به ۴ بخش مجزا تقسیم‌بندی نمود:

۱. **اعمال ورودی به مقدم‌ها:** در این مرحله ورودی‌های به مقدم‌ها داده می‌شود و درجه عضویت هریک از آنها بدست می‌آید.

۲. **اعمال عملگرهای فازی:** هنگامی که تعداد مقدم‌ها از یک بیشتر شود، باید عملگرهای فازی به‌کار گرفته شوند تا عددی به دست آید که نمایانگر حاصل مقدم‌ها بر آن قانون باشد. سپس این عدد در تابع خروجی به کار گرفته شود. این عدد، "عدد درستی" آن قانون نامیده می‌شود. از مهم‌ترین روابطی که در این قسمت وجود دارد، رابطه استلزام ممدانی و لارسن است. رابطه استلزام ممدانی

و لارسن به ترتیب از عملگرهای min و ضرب برای به دست آوردن عدد درستی هر قاعده استفاده می‌کنند.

$$R_{mamdani}(u,v) = \min[\mu_A(u), \mu_B(v)] \quad (۱۵-۲)$$

$$R_{larsen}(u,v) = \mu_A(u).\mu_B(v) \quad (۱۶-۲)$$

۳. **اعمال روش دلالت:** ورودی روش دلالت عدد درستی و خروجی آن یک مجموعه فازی خروجی است. قبل از اعمال روش دلالت، ذکر این نکته ضروری است که وزن هر قاعده، که توسط خبره یا هر روشی تعیین می شود، باید در عدد درستی ضرب شده، سپس به تابع خروجی اعمال شود. در صورت عدم وزن‌دار بودن روابط وزن تمامی آنها یک در نظر گرفته می‌شود.

۴. **اجتماع خروجی‌ها:** از آنجایی که تصمیم‌گیری بر اساس تمام قواعد در سیستم‌های استنتاج فازی صورت می‌گیرد، لذا برای تصمیم‌گیری باید قواعد را به روشی به‌توان ترکیب کرد. اجتماع در واقع روشی است که به واسطه آن تمام مجموعه‌های خروجی هر قاعده به یک مجموعه فازی واحد ترکیب می شود. ورودی فرآیند اجتماع، فهرستی از توابع خروجی است که به وسیله فرآیند دلالت برای هر قاعده بریده شده و خروجی آن یک مجموعه فازی برای خروجی است. روش‌های متفاوتی برای اجتماع گیری وجود دارد که مهم‌ترین آن‌ها ماکزیم‌گیری و جمع است که ما در موتور استنتاج فازی از روش جمع استفاده می‌کنیم. علت استفاده از روش جمع، این است که روش جمع مجموع قوانین را در نظر می گیرد، در صورتی که در روش ماکزیم، آن قاعده ای که مقدار ماکزیم را دارد، را در نظر گرفته و بقیه قوانین را نادیده می گیرد.

۲-۴-۲- سیستم استنتاج فازی محاسبه شباهت

در محاسبه شباهت به روش استنتاج فازی، آنگونه که ذکر شد با تعریف برخی ویژگی‌ها، سیستم استنتاج فازی وظیفه تصمیم‌گیری در قبال شباهت ورودی‌ها را خواهد داشت. این روش‌ها دارای سه جزء اصلی پیش‌پردازش، قطعه‌بندی و استخراج ویژگی هستند و در نهایت سیستم استنتاج فازی وظیفه ارزیابی دو محتوا را ایفا می‌کند. در این بخش به بررسی تک‌تک اجزا مخصوصاً بخش سیستم استنتاج فازی خواهیم پرداخت.

۲-۴-۲-۱- پیش‌پردازش و قطعه‌بندی

در این مرحله عملیاتی از قبیل یکسان‌سازی نگارشی و حذف ایست‌واژه‌ها انجام می‌شود. اما از آنجا که مقالات بررسی شده در این حوزه ورودی‌ها را از لحاظ معنایی و محتوایی نیز مورد بررسی قرار می‌دهند برخی از فرآیندها مانند ریشه‌یابی افعال نیز به این فرآیند افزوده می‌شود.

۲-۴-۲-۲- استخراج ویژگی

در این مرحله بخش‌های مرتبط با هم مقایسه می‌گردد. برخی ویژگی‌های استفاده شده در این مقاله در ذیل آورده شده است.

۱. **معیار F**: این معیار تعداد واژه‌های مشترک بین دو متن را از طریق تولید n-گرام‌ها (در سطح واژه و کاراکتر) را به ترتیب تقسیم بر طول متن اول و دوم کرده و سپس از این دو نسبت میانگین می‌گیرد [۳۷].

۲. **ســـنجش پرش-گرام:** این ســـنجه همانند n-گرام اســـت با این تفاوت که می‌تواند برای تولید گرام‌ها در سطح واژه پرشی به اندازه n داشته باشد که اندازه n عدد سه و چهار انتخاب شده است. این نوع سنجش عبارات چند واژه‌ای مشترک در هر بخش را در نظر می‌گیرد [۳۳].

۲-۴-۲-۳- استنتاج فازی

پس از تعریف سنجه‌ها و ویژگی‌های مناسب، از آنجا که سنجش شباهت بین دو متن عملی است که با توجه به دانش افراد خبره که از منظرهای گوناگون به شباهت نگاه می‌کنند، متفاوت و مبهم است و برای درک درست از این عمل، بهتر است آن را به صورت فازی مدل کرده تا قادر به برخورد با ابهام و نامعلوم بودن سنجش بود. در این بخش در ابتدا برخی از متغیرهای فازی تعریف می‌شوند. تعریف این متغیرها به این صورت خواهد بود که به ازای هر ویژگی یک متغیر فازی خاص آن ویژگی تعریف خواهد شد. از آنجا که ارزیابی شباهت بین دو متن به صورت فازی تعریف شده است، معیارهایی که در ورودی سیستم‌های استنتاج محاسبه می‌شوند نیز به صورت فازی مدل می‌شوند.

پس از تعریف دقیق متغیرهای فازی سیستم استنتاج قواعد فازی با استفاده از خبرگان و متخصصان استخراج می‌شود. این قواعد در نهایت موتور اصلی سیستم استنتاج فازی محاسبه شباهت را تشکیل خواهند داد. تعریف متغیرها تا جایی باید صورت پذیرد که تقریبا تمامی حالات محتمل سیستم را تشکیل دهد. هر چه این قواعد بیشتر باشد سیستم کارایی بهتری در مواجهه با ورودی‌های متفاوت خواهد داشت [۳۵].

۲-۵- بررسی کارهای مرتبط

تاکنون برخی از تکنیک‌های استفاده شده که در مقالات فارسی و انگلیسی بسیار پرکاربرد بوده‌اند را بررسی کردیم. به طور کلی می‌توان گفت سیستم‌های استنتاج فازی، اثر انگشت، مدل فضای بردار و خوشه‌بندی از

مهم‌ترین این تکنیک‌ها در زمینه محاسبه شباهت لغوی به شمار می‌روند. در این بخش مروری خواهیم داشت بر کارهای مرتبط دیگر که در سال‌های اخیر انجام شده است.

لانگ و همکاران [۳۸] در روشی دو مرحله‌ای از طریق الگوریتم اثر انگشت و روش خوشه‌بندی به کشف سرقت ادبی پرداختند. این روش، نوعی روش آماری محسوب می‌شود که به زبان وابسته نیست. در گام اول تمامی اثر انگشت‌ها در یک فضای برداری قرار می‌گیرند. در این مرحله اگر بیشتر از یک دنباله مشابه در بین اثر انگشت‌ها وجود داشته باشد به این معنا است که یک سری اثر انگشت پشت سر هم به سرقت رفته‌اند. در گام دوم محل این مشابهت‌ها مشخص می‌گردد. در مرحله اول از تکنیک Winnowing و روش پنجره کشویی برای استخراج اثر انگشت‌های متن استفاده شده است. در گام دوم با خوشه‌بندی سری‌های متوالی ایجاد شده از اثر انگشت‌ها، بخش‌های مشکوک شناسایی می‌شود.

مشابه روش لانگ را الظهرانی و سلیم [۳۹]، با در نظر گر جمله به عنوان بخش‌های جداکننده به قابل سرقت متون انجام داده‌اند. در این روش از تکنیک اثر انگشت و ایجاد ۴ گرام‌ها برای تولید ویژگی استفاده شده است. در این روش جملات به شیوه‌ای از یکدیگر جدا می‌شوند. در نهایت ۱۵ سند با یکدیگر مقایسه می‌شوند. در این روش دو تکنیک اثر انگشت و روش بازیابی فازی نیز با هم مقایسه می‌گردند.

اکبال و همکاران [۴۰] روشی مبتنی بر مدل فضای بردار ارائه کردند. پس از پیش‌پردازش، تمای کلمات در فضای برداری (VSM) مدل می‌شوند. در این مرحله هر واژه با معیار TF-IDF[۱] وزن‌دهی می‌شود. سپس هر سند منبع و مشکوک به صورت بردار وزنی نمایش داده می‌شود و تشابه میان این دو سند با محاسبه معیار کسینوسی صورت می‌پذیرد. در مرحله بعدی بردار n-گرام اسناد ایجاد می‌شود. سپس این n-گرام‌ها توسط

[۱] **Term frequency – Inverse document frequency**

ضریب جاکارد مقایسه می‌شوند و ضریب تشابه آنها استخراج می‌گردد. اگر ضریب تشابه بیش از حد آستانه باشد طی عملیاتی مبتنی بر تکنیک‌های گراف قطعات سرقت شده استخراج می‌شود.

محمودی و همکاران [۴۱] به کشف سرقت ادبی در زبان فارسی پرداخته‌اند. در این پژوهش نیز یک روش سرقت ادبی در محیط تک‌زبانه و مستقل از زبان و مبتنی بر محتوا ارائه شده است. برای کشف شباهت دو متن از یک فرهنگ لغت برای در نظر گرفتن مترادف‌ها نیز استفاده شده است. برای مقایسه شباهت دو سری از کلمات تولید می‌شود که از ۲ و ۳ گرام استفاده شده است. سپس با معیار جاکارد و در نظر گفتن حد آستانه برای شباهت، تشابه بین دو سند محاسبه می‌شود. پس از آن نسبت واریانس به میانگین کاندیدهای شباهت محاسبه می‌شود. این ضریب پراکندگی مشخص می‌کند گرام‌های مشابه چقدر به یکدیگر نزدیک بوده‌اند. اگر ضریب پراکندگی عدد بالایی باشد، نمی‌توان دقت را در بین جملات محاسبه کرد، بنابریان ممکن است دو متن کلمات کلیدی مشترک زیادی داشته باشند و سرقتی رخ نداده باشد.

ممتاز و همکاران [۴۲] روشی برای تشخیص سرقت علمی اسناد فارسی با رویکرد گراف ارائه کرده‌اند. در این روش از ساختار گراف و روش تکرار در مشابهت گره‌های درون گراف برای یافتن تشابه دو سند استفاده شده است. برای تشخیص سرقت یک حد آستانه برای روش تکرار در مشابهت گره‌ها در نظر گرفته می‌شود. در نهایت ارزیابی روی پیکره PAN فارسی [۳] صورت می‌گیرد.

آهنگربان و منتظر [۴۳] روشی مبتنی بر سیستم نروفازی برای کشف شباهت بین اسناد ارائه کردند. در این روش از قواعد فازی برای استنتاج برای ویژگی‌ها استفاده شده است. برای یادگیری از داده‌های آموزشی از سیستم نروفازی استفاده شده است. در این روش موجودیت کلمات به دو بخش عمومی و تخصصی شکانده می‌شود. هر یک از این دسته واژگان با یکدیگر مقایسه شده. برای ارزیابی سیستم نیز از پیکره همیافت استفاده شده است.

پاول و جمال [۴۴] روشی برای سرقت ادبی ارائه کردند که از تکنیک ربته بندی جملات جهت کشف شباهت عبارات استفاده می‌کند. این پژوهش از روش برچسب گذاری معنایی نقش هر واژه از جمله استفاده می‌کند. در بخش پیش‌پردازش کلمات معنادار نگاه داشته می‌شوند و باقی واژه‌ها حذف می‌گردند. سپس در مرحله‌ای تحت عنوان بازیابی سند، یک زیر مجموعه از سند اصلی ساخته می‌شود که شامل مقایسه جامع بین سند مشکوک و تعداد زیادی سند اصلی است. مقایسه این سندها نیز از طریق n-گرام‌های با طول ۲، ۳ و ۴ صورت می‌گیرد. سپس برای رتبه دادن به جملات در سند اصلی و مشکوک، از یک تکنیک رتبه‌بندی برای محاسبه رتبه هر جفت جملات اصلی و مشکوک استفاده می‌شود. برای این کار نیز از شباهت کسینوسی استفاده می‌شود. در این تکنیک بسیار از جملات غیرضروری کنار گذاشته می‌شود. در این تکنیک جملات با بیشترین شباهت به دست می‌آید. همچنین در این روش از تکنیک SRL که فرآیند ارتباط معنایی بین فعل و اجزای تشکیل دهنده جمله را کشف می‌کند، استفاده شده است.

۲–۶- سیستم‌های موجود محاسبه مشابهت

تاکنون سیستم‌های بیشماری برای محاسبه مشابهت بین دو عبارت و همچنین تشخیص سرقت علمی-ادبی و کپی‌برداری‌های الکترونیکی نگارش شده است. در این قسمت به معرفی برخی از این سیستم‌ها می‌پردازیم.

سیستم Turnitin[۱] و Plagirism[۲] یکی از معروف‌ترین این سیستم‌ها می‌باشد. نحوه کار این سیستم‌ها مبتنی بر الگوریتم اثر انگشت می‌باشد. بدین صورت که مجموعه‌ای از اثر انگشت‌های بدست آمده از اسناد مختلف را در پایگاه داده خود دارد. در این سیستم، اسناد دریافت می‌شود و پس از بررسی مشخص می‌شود که چند درصد از اسناد کپی‌برداری شده است و اینکه این اسناد چقدر به یکدیگر شبیه هستند.

---

[۱] **Turnitin.com**

[۲] **Plagiarism.org**

یکی دیگر از این نوع سیستم‌ها Edutie[1] می‌باشد که در آن با قرار دادن اسناد، با استفاده از الگوریتمی که مشابه با winnowing است، به بررسی اسناد می‌پردازد. مزیت این سیستم ویژگی‌های واسط کاربری و مدیریت کارای اسناد می‌باشد.

سیستم JPlag نیز برای مشابهت‌های کدهای برنامه‌نویسی ایجاد شده است. این سیستم بر روی مشابهت کدهای برنامه‌نویسی به زبان‌های مختلف تمرکز دارد. این سیستم از زبان‌های برنامه‌نویسی مختلفی پشتیبانی می‌کند. این سیستم نیز مانند سیستم‌های پیشین تحت وب است.

از جمله سیستم‌هایی که به صورت نرم‌افزار قابل ارائه هستند می‌توان به EVE2 اشاره کرد. این سیستم یک نرم‌افزار می‌باشد که بر روی رایانه‌های شخصی قابل نصب است و از اسناد با فرمت‌های Word و Plain Text پشتیبانی می‌کند. این نرم‌افزار توسط شرکتی کانادایی ارائه شده که نسخه آزمایشی ۱۵ روزه آن روی سایت شرکت سازنده موجود است.

از نمونه‌های داخلی کشف سرقت ادبی می‌توان به سیستم کشف سرقت سمیم نور اشاره کرد. این سیستم پیکره فارسی مشتمل بر چهار هزار سند مرجع و سند مشکوک به سرقت را برای ارزیابی سیستم‌های کشف سرقت ارائه کرده است.

۲-۷- چالش های زبان فارسی

برخلاف زبان انگلیسی که در آن هم حروف و هم لغات کاملاً متمایز از یکدیگر هستند، در زبان فارسی پیوستگی میان برخی علائم یا لغات وجود دارد. علاوه بر آن تنوع نگارشی در کلمات نیز موجود می‌باشد. به عنوان مثال ریشه‌یابی فعل در یک لغت بهم پیوسته در زبان فارسی چالش‌های خاص خود را دارد. به عنوان

[1] **Edutie.com**

مثال در یک لغت بهم پیوسته هم بن فعل، شناسه، علامت زبان فعل و حتی شناسه‌های مفعولی می‌توان داشت که کار پردازش لغات را پیچیده‌تر می‌نماید. به گونه که نمی‌توان از دانش، تجربه و سایر نرم افزارها و کتابخانه‌های موجود بر روی زبان‌های برنامه‌نویسی در این زمینه استفاده نمود. و همچنین تولید نرم افزاری که قادر به حل تمامی این پیچیدگی‌ها باشد، فرآیندی زمان‌بر و مستلزم تلاش‌های فراوان است. تفاوت‌های ذاتی زبان‌های گسسته مانند زبان انگلیسی با زبان‌های مانند زبان فارسی که با یکدیگر تفاوت‌های بنیادین دارند، منجر شده که ادعای اعمال تغییرات در ساختار یک نرم افزار انگلیسی و بدست آوردن نتایج خوب برای زبان فارسی لزوماً امکان پذیر نباشد و مستلزم آزمایش‌های فراوان برای اثبات صحت آن خواهد بود. در ادامه به برخی از مشکلات موجود در ربان فارسی اشاره می‌نماییم که در کار انجام شده در این پایان نامه نیز با آنها برخورد داشته‌ایم.

۱) **نبود نکات گرامری تعریف شده:** همانند آنچه که در زبان انگلیسی وجود دارد. این امر در ریشه‌یابی و پیش پردازش متن تاثیر خواهد گذاشت.

۲) **فاصله و نیم‌فاصله:** یکی از مهم‌ترین چالش‌های در ارتباط با متن فارسی وجود فاصله و نیم‌فاصله بین پسوندها، پیشوندها و ریشه است. همچنین در زبان فارسی دسته‌ای از فعل‌ها که به آن‌ها فعل مرکب گفته می‌شود، دارای چند بخش هستند، با توجه به اینکه در دستور زبان فارسی هر فعل مرکب یک کلمه است، باید بین بخش‌های آن نیم‌فاصله وجود داشته باشد. عدم رعایت این نکته در بسیاری از متون عملکرد سیستم را با مشکل مواجه می‌کند.

۳) **وجود برخی کاراکترها با چند یونیکد:** در زبان فارسی کاراکترهایی وجود دارند که دارای چندین فرم یونیکد می‌باشند. این مشکل به دلیل مشترک بودن زبان فارسی با زبان‌های دیگر نظیر عربی

و کردی است. به عنوان مثال "می‌روم" و "می‌روم" هر دو یک ریشه با عنوان "رو" دارند که به دلیل متفاوت بودن کاراکتر "ی" فقط کلمه "می‌روم " با ریشه خود منطبق است.

۴) **کاراکترهای غیر متداول:** در زبان فارسی کاراکترهایی وجود دارند که در بسیاری از موارد در متون نوشته نمی‌شوند و این به آن دلیل است که کلمه‌ها در بیشتر موارد مربوط به زبان عربی هستند که شامل این کاراکترهای غیرمتعارف می‌باشد.

۵) **وجود صورت عامیانه کلمات در متن:** در زبان فارسی در بسیاری از موارد کلمه در فرم عامیانه کلمه متفاوت از فرم نوشتاری آن بیان می‌شود. در بسیاری از متون این فرم عامیانه به زبان نوشتاری هم راه می‌یابد و در فرآیند پردازش متن مشکل ایجاد می‌کند. این نوع استفاده از کلمات به خصوص با رشد و توسعه شبکه‌های اجتماعی در سال‌های اخیر بیشتر شده‌است. به عنوان مثال "خونه" حالت عامیانه کلمه "خانه" است که توسط ریشه‌یاب‌ها قابل تشخیص نیست.

۶) **چند معنایی:** در زبان فارسی کلماتی وجود دارند که با وجود شکل یکسان دارای معانی متفاوت هستند. نقطه تمایز این کلمه‌ها مصوت‌هایی است که معمولا در نوشته‌ها از درج آن‌ها صرفه نظر می‌شود. به عنوان مثال "سَر" و "سِر" دوکلمه کاملاً متفاوت از نظر معنایی هستند که در حوزه شباهت تشخیص آنها از یکدیگر بر حسب معنا کار پیچیده‌ای است.

۷) **استثنأها در قوانین اضافه کردن پیشوند و پسوند در موارد خاص:** اگر چه برای ساختن کلمات از ریشه‌های موجود توسط اضافه کردن پیشوندها و پسوندها قوانین ثابت و مشخص وجود دارد، اما با توجه به نحوه تلفّظ حروف در بعضی موارد این قوانین دست‌خوش تغییرات قرار می‌گیرند. به عنوان مثال قانون کلی برای ساخت فعل اول شخص مفرد اضافه کردن "م"به بن مضارع فعل است. در مورد مصدر "جو" به معنی جست و جو کردن، این قانون با توجه به اینکه "و" یک حرف

صدا دار یا مصوت است، با اندکی تغییر اعمال می‌شود. به این صورت که قبل از "م" حرف "ی" به کلمه اضافه می‌شود تا کلمه قابل خواندن باشد و نتیجه در این مورد "جویَم" خواهد بود. این مشکل زمانی پیچیده‌تر می‌شود که "جو" با توجه به این که "و" صامت تلفّظ شود یا مصوت دو مفهوم متفاوت خواهد داشت.

۸) **وجود لغات ترکیب جزئی:** "مانند آب سرد کن" و "سکه بهار آزادی".

۹) **استفاده از کلمات انگلیسی به صوت‌های مختلف:** مانند "سورس" و "source".

۱۰)**استفاده از برخی کلمات به جای یکدیگر برای ساده‌نویسی:** مانند "آ" و "ا" به جای یکدیگر در کلمات "فرآیند" و "فرایند".

۱۱)**تنوع استفاده از "می" چسبان و غیرچسبان:** همانند کلمات "می‌تواند" و "میتواند".

۱۲)**تنوع به کاربردن "ها" چسبان و غیرچسبان:** مانند "آن‌ها" و "آنها".

۱۳)**تنوع نگارش "ی" اضافه در کلمات مختوم به "ه":** مانند "خانه‌ی سبز" و "خانه سبز".

۱۴)**فقدان هر نوع منبع داده‌ای و پیکره مناسب:** مانند یک گرامر کامل، یک مجموعه از جملات تجزیه شده و یا آمارهای ارزشمند از کاربرد لغات در جملات مختلف و جایگاه‌های مختلف در جمله.

۱۵)**نبود حروف بزرگ و کوچک همانند حروف زبان انگلیسی برای تشخیص اسامی خاص:** مانند نام‌ها، روزهای هفته، ماه و مناطق جغرافیایی.

۱۶)**فقدان کتابخانه‌های برنامه‌نویسی پردازش زبان‌های طبیعی مختص زبان فارسی:** برای زبان انگلیسی کتابخانه‌های فراوانی برای پردازش و پیش‌پردازش متون مورد استفاده قرار می‌گیرند که

تقریباً با تمام زبان‌های برنامه‌نویسی سازگار هستند. اما متاسفانه برای زبان فارسی چنین کتابخانه‌هایی وجود ندارد و پژوهشگران برخی فرآیندها را به اجبار تکرار می‌کنند.

۲-۸- نتیجه‌گیری

در این فصل ضمن تعریف مفاهیم پایه در محاسبه شباهت بین اسناد، انواع روش‌های مورد استفاده در این زمینه معرفی شد. روش‌های مبتنی بر رشته و مستقل از زبان مورد ارزیابی قرار گرفت. روش‌های مبتنی بر انگشت‌نگاری جزو مهم‌ترین این روش‌ها بودند که به تفصیل بررسی شدند. این روش‌ها به عنوان روش‌های پایه روش پیشنهادی قرار گرفته‌اند. همچنین مروری بر تحقیقات مرتبط با روش پیشنهادی انجام شد. سیستم استنتاج فازی که به عنوان تصمیم‌گیرنده در شرایط عدم قطعیت شناخته می‌شود تشریح شد و کارهای مرتبط بر مبنای این سیستم نیز بیان شد. در پایان نیز به چالش‌های اصلی پردازش متن در زبان فارسی که در محاسبه شباهت موجب کاهش کارایی سیستم می‌شوند مطرح گردید.

فصل سوم

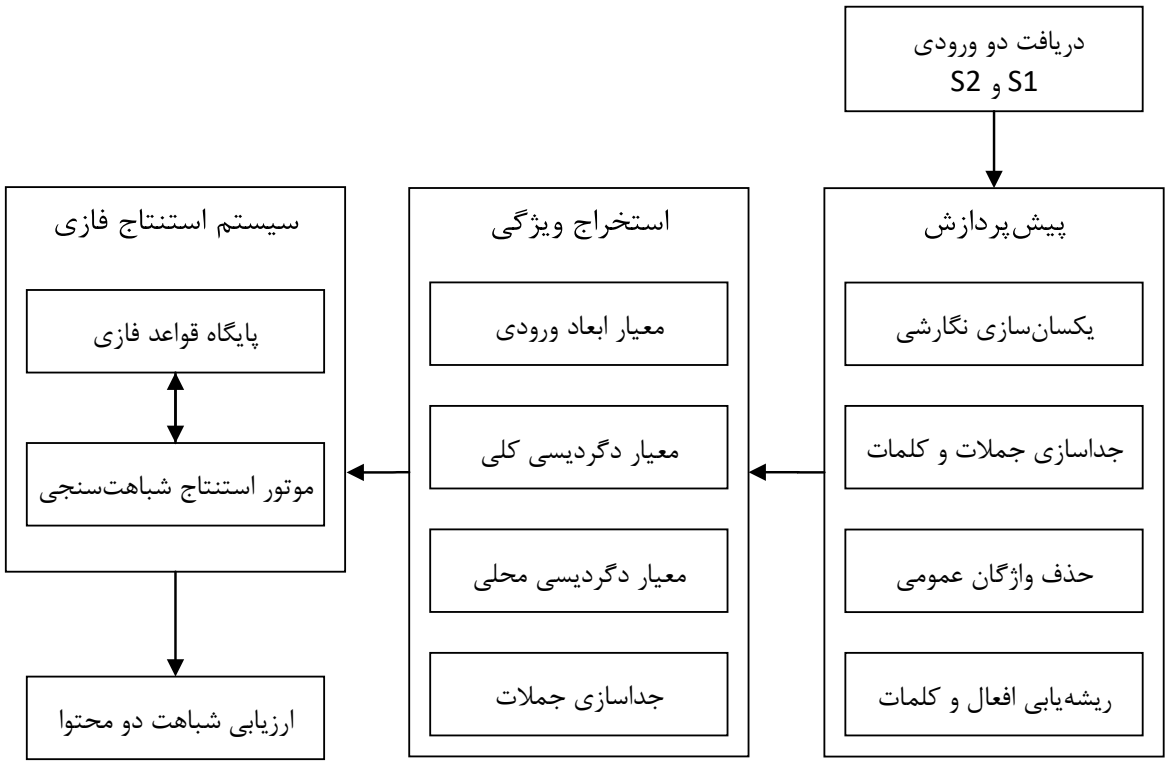
روش پیشنهادی

۳- روش پیشنهادی

در فصل گذشته، روش‌های کلی محاسبه شباهت میان دو رشته ورودی از دیدگاه‌های متفاوت بررسی شد و به طور ضمنی با بررسی کارهای گذشته به محدودیت‌ها و نقاط ضعف روش‌های موجود محاسبه شباهت بر اساس رشته اشاره شد. در این فصل، با ارائه یک معیار محاسبه شباهت و توسعه این معیار، سعی شده است که عملکرد سیستم محاسبه تشابه لغوی بین عبارات بهبود یابد. در نهایت نتایج حاصل از اعمال این روش‌ها، خطای خروجی را کاهش داده و موجب دقت بیشر و نزدیک‌تر به نمونه‌های آزمایشی پیکره مورد ارزیابی شده است. از این رو در این فصل به طراحی چارچوب و پیاده‌سازی این سیستم می‌پردازیم.

۳-۱- معماری روش پیشنهادی

در این بخش ابتدا یک بلوک دیاگرام از روش پیشنهادی برای سیستم مقایسه عبارات ارائه داده‌ایم و سپس به صورت مفصل مراحل مختلف آن تشریح می‌شود. شکل ۳-۱، معماری پیشنهادی را با جزئیات لازم نمایش می‌دهد. همانطور که ملاحظه می‌کنید این معماری دارای سه بخش اصلی پیش‌پردازش، استخراج ویژگی و سیستم استنتاج فازی است.



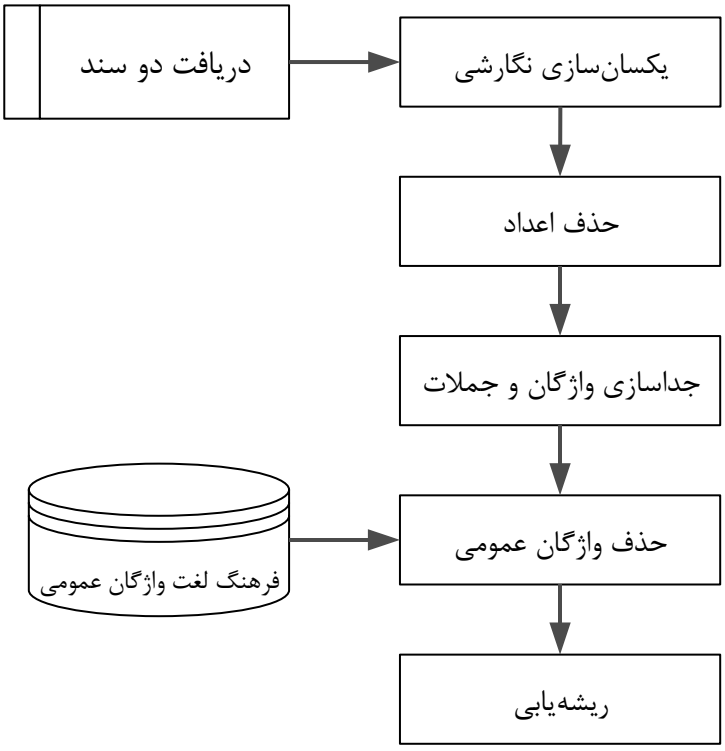
شکل ۳-۱ معماری سیستم پیشنهادی

۳-۲- دریافت عبارات

در اولین مرحله سیستم دو ورودی برای محاسبه شباهت دریافت می‌کند. این دو عبارت از چند جهت می‌توانند دسته‌بندی شوند که در ادامه در فاز انتخاب و استخراج ویژگی‌ها، این ورودی‌ها به سه دسته عبارات کوتاه در مقیاس کلمه‌ها و جملات، متون متوسط و متون طولانی تقسیم می‌شوند. دو عبارت ورودی می‌توانند به هر زبان دلخواهی به عنوان ورودی انتخاب شوند. البته اگر دو عبارت دارای زبان‌های مغایر یکدیگر باشند، خروجی دارای بیشترین فاصله ممکن خواهد بود و به این معنا است که دو ورودی هیچ وجه شباهتی ندارند.

۳-۳- پیش‌پردازش

عبارات ورودی پیش از ورود به مرحله انتخاب ویژگی باید پیش‌پردازش شوند. پیش‌پردازش شامل انجام فعالیت‌هایی بر روی متن است که موجب بهبود نتایج الگوریتم‌های محاسبه شباهت می‌شود. همچنین این فعالیت‌ها باعث افزایش دقت و کاهش زمان محاسبه شباهت نیز می‌گردد. شکل ۳-۲ مراحل پیش‌پردازش را نمایش می‌دهد.



شکل ۳-۲ مراحل عملیات پیش‌پردازش

۳-۳-۱- یکسان‌سازی نگارشی

این مرحله به جهت یکسان‌سازی نگارشی زبان فارسی صورت می‌گیرد. ممکن است نویسه‌های به کار رفته در دو واژه یکسان، باهم متفاوت باشند. این امر باعث می‌شود که هنگام شمردن واژه‌ها، دو واژه یکسان با

املاهای متفاوت به عنوان دو واژه مختلف در نظر گرفته شوند. برای جلوگیری از بروز این مشکل، نیازمند یکسان‌سازی متون ورودی هستیم.

در مرحله یکسان‌سازی ابتدا حروف مشابه را دسته‌بندی کرده و با استفاده از کد اسکی آنها، یک حرف را به جای حروف مشابه در تمامی کلمات جایگزین کرده. بنابراین یک کلمه با دو حرف مشابه، اما با کد اسکی متفاوت نخواهیم داشت. عمده حروفی که دارای کد اسکی هستند حروف ی و ک هستند که در دو زبان فارسی و عربی کدهای متفاوتی دارند. همچنین پیشوند "می" و پسوند "ها" در ابتدا و انتهای واژه‌ها، ممکن است به سه صورت مختلف زیر دیده شوند:

بدون فاصله	چسبان جدا	با فاصله جدا
کتابها	کتاب‌ها	کتاب ها
میرود	می‌رود	می رود

و یا اینکه برخی واژه‌ها مانند «مسئول» و «مجموعه‌ی» می‌توانند به صورت‌های زیر در متون فارسی دیده شوند.

مسئول	مسوول	مسؤول
مجموعه‌ی	مجموع ی	مجموعه

برای نرمالسازی متون انگلیسی کتابخانه‌های متفاوتی ارائه شده‌اند که معروف‌ترین آنها کتابخانه پردازش زبان‌های طبیعی دانشگاه استنفورد[۱] است. برای زبان فارسی با وجود آزمایشگاه‌های مختلف در زمینه پردازش زبان‌های طبیعی مانند آزمایشگاه پردازش متن و زبان طبیعی دانشگاه تهران[۲] و آزمایشگاه فناوری وب

[۱] **Stanford Natural Language Processing Lib**

[۲] **ece.ut.ac.ir/lab/nlp**

دانشگاه فردوسی[1]، اما همچنان برای انتشار کتابخانه رسمی پردازش متن به زبان فارسی متولی خاصی در این حوزه وجود ندارد. با این وجود برای پیش‌پردازش‌های زبان فارسی در محاسبه شباهت، ما از کتابخانه متن باز هضم استفاده نمودیم. این کتابخانه که برای زبان‌های جاوا و پایتون کتابخانه رسمی ارائه داده‌است، توابع پیش‌پردازش زبان فارسی را در خود دارد. برای پیش‌پردازش اولیه متون ورودی از تابع normalizer در این کتابخانه جهت نرمال‌سازی متون ورودی استفاده شده‌است که نمونه‌ای از آن در ذیل آورده شده است:

```
>> normalizer = Normalizer()
>> normalizer.normalize( ' اصلاح نویسه ها و استفاده از نیم‌فاصله پردازش را آسان می کند ' )
>> Output : ' اصلاح نویسه‌ها و استفاده از نیم‌فاصله پردازش را آسان می‌کند '
```

در مورد زبان انگلیسی نیز، نیاز به یک سری پیش‌پردازش داریم. در این بخش بایستی تمامی حروف را به حروف کوچک تبدیل کرده و سپس آنها را در مراحل بعد مورد بررسی قرار دهیم.

۲-۳-۳- حذف اعداد و متغیرها

برای حذف اعداد از روشی به نام توکن‌سازی استفاده می‌شود. این روش، روش بسیار مناسبی در تشخیص تشابهات بخصوص در برنامه‌های کامپیوتری است. الگوریتم توکن‌سازی عناصر کد برنامه را با توکن‌های واحدی جایگزین می‌کند [۴۵]. برای مثال هر متغیر[2] با توکن $< ID >$ جایگزین می‌شود. هر مقدار عددی نیز با $< Value >$ جایگزین می‌شود. حال اگر یک برنامه کامپیوتری شامل عبارتی به صورت $a = b + 4$ باشد، این عبارت با رشته $< ID > = < ID > + < Value >$ جایگزین می‌شود. بنابراین اگر در

[1] **wtlab.um.ac.ir**

[2] **variable**

برنامه‌های کامپیوتری کاربر نام متغیری را تغییر دهد، تغییری در ترجمه خط ایجاد نمی‌شود. این نوع جایگزینی باعث می‌شود اعداد به طور کلی از ویژگی‌های متون حذف نگردند. معمولا در سرقت یک متن ممکن است اعداد تغییر کند. این تغییر تاثیری در اصل اثر و ترتیب آن عدد به عنوان ویژگی نخواهد داشت.

۳-۳-۳- حذف واژه‌های عمومی

در این مرحله واژه‌های عمومی از قسمت‌های مختلف متن حذف می‌گردند. این کار عملیات بررسی را سریع‌تر خواهد کرد. همچنین قسمت مهم محتوا را نگاه خواهد داشت و قدرت تصمیم‌گیری در مورد شباهت متون افزایش می‌یابد. واژه‌های عمومی، واژه‌هایی هستند که اهمیت آنها در جمله ناچیز است. واژه‌های عمومی از واژه‌های پرتکرار و یا ایست‌واژه‌ها[۱] مجزا است. در سال ۲۰۰۹ سنجی و آرمیده از دانشگاه فردوسی مشهد، یک مجموعه متون در زمینه‌های روانشناسی، آموزشی، علوم اطلاعاتی و یک لیست ۹۲۰ کلمه‌ای تهیه کردند [۴۶]. این لیست با استفاده از نسخه برخط روزنامه همشهری، لغت‌نامه دهخدا و فرهنگ سخن تهییه شده است. این واژه‌ها را در مرحله پیش‌پردازش از متون ورودی حذف می‌کنیم. اما برای مقایسه تاثیرگذاری این روش، متون اصلی را نیز برای انجام آزمایشات، اصل اثر همراه واژه‌های عمومی را نیز نگاه خواهیم داشت.

۴-۳-۳- جداسازی واژگان

پس از نرمالسازی متون ورودی و ریشه‌یابی کلمات، متون ورودی به وسیله جداسازی واژگان به توکن تبدیل می‌شوند. شناسایی و جداسازی واژه‌ها از طریق بررسی علائم قابل انجام است. این علائم برای زبان فارسی عبارت‌اند زا: "،"، "؟"، ":"، "؛"، "!"، "<"، ">"، "-"، "_"، "/"، "("، ")".

[۱] **Stop Words**

البته برخی از واژه‌های مرکب در این روش شکسته شده، به گونه‌ای که تبدیل این واژه‌های مرکب (مانند "سکه بهار آزادی") به توکن‌های جدا از هم، معنای آن واژه را تغییر می‌دهد. از این جهت لازم است جدای از تبدیل واژه‌ها به توکن‌ها از الگوریتم‌هایی همانند n-گرام نیز برای در نظر گرفتن این نوع عبارات استفاده شود.

۵-۳-۳- جداسازی جملات

علاوه بر تبدیل متون به توکن‌ها و استخراج کلمات، بر اساس ویژگی‌های در نظر گرفته شده در این پژوهش بر اساس برخی از ویژگی‌ها، نیاز است متون به جملات جدا از هم نیز شکسته شوند. در زبان فارسی جملات به علائمی مانند "."، "؟" و "!" ختم می‌شوند. از این رو جملات نیز به توکن‌های شامل کلمات تبدیل می‌شوند. برای جداسازی جملات ما از کتابخانه پردازش زبان هضم استفاده نموده‌ایم.

۶-۳-۳- ریشه‌یابی کلمات و افعال

ریشه‌یابی کلمات[۱] و ریشه‌یابی فعل‌ها[۲] را از این جهت جزو پیش‌پردازش زبان در نظر می‌گیریم که می‌توانند به بهبود الگوریتم‌ها کمک بسزایی کنند. این بخش برای هر زبان متفاوت خواهد بود. به عنوان مثال برای زبان انگلیسی از کتابخانه‌ها و فرهنگ لغات[۳] خاص آن زبان برای ریشه‌یابی کلمات استفاده می‌شود. بنابراین این بخش به جهت بهبود کارایی الگوریتم‌های محاسبه شباهت مورد استفاده قرار می‌گیرد. ما در این بخش نیز از کتابخانه هضم برای ریشه‌یابی کلمات و عبارات استفاده کرده‌ایم که نمونه‌ای از این فراخوانی‌ها نمایش داده شده است.

---

[۱] **Stemming**

[۲] **Lemmatization**

[۳] **Dictionary**

```
>>> stemmer = Stemmer()
>>> stemmer.stem('کتاب‌ها')
'کتاب'
>>> lemmatizer = Lemmatizer()
>>> lemmatizer.lemmatize('می‌روم')
'رفت#رو'
```

۳-۴- انتخاب و استخراج ویژگی‌ها

در این مرحله برخی از ویژگی‌ها و معیارهایی که برای محاسبه شباهت بین عبارت مورد ارزیابی قرار داده‌ایم را معرفی خواهیم نمود. برخی از این ویژگی‌ها بر اساس کارهای مرتبط به صورت پایه‌ای وجود داشته و ما در این قسمت سعی داریم تا با استفاده از آنها به بهبود الگوریتم دست یابیم.

۳-۴-۱- مقیاس دو ورودی

یکی از ویژگی‌های مد نظر جهت تصمیم‌گیری در قبال چگونگی عملکرد الگوریتم با ورودی‌ها مقیاس ورودی‌ها است. در این مرحله دو رشته دریافت شده از لحاظ ابعاد و طول رشته ارزیابی و طبقه‌بندی می‌شوند. دو عبارت در یکی از سه طبقه کوچک، متوسط و بزرگ قرار خواهند گرفت. تقسیم‌بندی مقیاس ورودی‌ها از این جهت پر اهمیت است که سیستم بر اساس طول رشته ورودی در مورد میزان شباهت آنها تصمیم‌گیری خواهد کرد. بنابر ابعاد ورودی الگوریتم‌های متفاوتی برای محاسبه شباهت به کار می‌روند که در بخش‌های بعدی مفصلاً در مورد آنها بحث خواهد شد.

معیارهای تقسیم‌بندی رشته بر اساس طول کلمات است. تقسیم‌بندی کلمات به صورت زیر خواهد بود.

جدول ۳-۱ مقیاس دو ورودی بر اساس تعداد کلمات هر ورودی

دسته‌بندی	اندازه ورودی (کلمه)
کوچک	۰ < ورودی ≤ ۲۰۰
متوسط	۲۰۰ < ورودی ≤ ۱۴۰۰
بزرگ	۱۴۰۰ < ورودی < ∞

۳-۴-۲- دگردیسی کلی[۱]

دگردیسی‌های کلی شامل تغییراتی است که در سرتاسر دو عبارت ورودی اتفاق می‌افتد. این دگردیسی‌ها می‌توانند ناشی از حذف، افزودن و یا جابجایی کاراکترها کلمات باشند. در بحث سرقت علمی-ادبی ناشیانه، معمولا کلیّت اسناد شامل جملات و کلمات کلیدی کپی‌برداری می‌شوند و تغییرات بسیار کمی در برخی قسمت‌ها اعمال می‌شود. به گونه‌ای که توسط یک معیار با پیچیدگی الگوریتمی کم می‌توان این نوع شباهت را تشخیص داد. در این بخش از عبارات ورودی در چند بخش مختلف هیستوگرام کاراکترها را استخراج نموده و این هیستوگرام‌ها با یکدیگر مقایسه می‌شوند. پس از عملیات پیش‌پردازش و جداسازی جملات از هر جمله به صورت مستقل هیستوگرام گرفته می‌شود. قابل ذکر است در صورت کوچک بودن ورودی، از تمام آن به صورت یکجا هیستوگرام گرفته می‌شود. روند مقایسه هیستوگرام کاراکترهای جملات به این صورت است که پس از تشکیل آنها، در نهایت درایه‌های دو هیستوگرام با یکدیگر به تفاضل می‌رسند و فاصله اقلیدسی دو هیستوگرام محاسبه می‌شود.

عدد خروجی تفاضل هیستوگرام‌ها بین صفر و بینهایت خواهد بود. از این رو برای نرمالسازی عدد خروجی و قرارگیری آن بین صفر و یک که نشانگر فاصله کلی دو هیستوگرام است، نیاز است در خروجی تفاضل

[۱] **Global Deformations**

تغییراتی اعمال شود. به همین جهت بدترین حالت ممکن که دو هیستوگرام می‌توانند نسبت به یکدیگر داشته باشند را در نظر می‌گیریم. به این صورت می‌توان نتیجه‌گیری کرد که بدترین حالت ممکن، زمانی اتفاق می‌افتد که یک کاراکتر در یک عبارت تکرار شده و در عبارت دیگر تکرار نشده باشد. بنابراین از هر دو هیستوگرام مورد مقایسه یک هیستوگرام خروجی خواهیم ساخت که مقدار فرکانس کاراکترهای آن برابر ماکزیمم تکرار کاراکترهای دو عبارت ورودی است. به عنوان مثال اگر حرف "ک" در عبارت اول دو بار و در عبارت دوم ۵ بار تکرار شده باشد، "ک" در هیستوگرام خروجی فرکانس مقدار ۵ را خواهد داشت.

در نهایت از هیستوگرام سوم نیز با هیستوگرام درایه صفر یک فاصله اقلیدسی گرفته می‌شود که عدد خروجی نماد بدترین حالت ممکن است. خروجی از تقسیم فاصله اقلیدسی دو هیستوگرام عبارات بر مقدار بدترین حالت ایجاد خواهد شد. این خروجی در نهایت مقداری بین صفر و یک دارد. رابطه (۱-۳) بیانگر این ویژگی می‌باشد.

$$GlobalDeformation(Hist_1, Hist_2) = \frac{\left(\sum_{i=1}^{chn} \left(Hist_{1i} - Hist_{2i}\right)^2\right)^{\frac{1}{2}}}{\left(\sum_{i=1}^{chn} \left(Hist_{worst_i} - 0\right)^2\right)^{\frac{1}{2}}} \qquad (۱-۳)$$

در رابطه فوق chn تعداد حروف مجاز در زبان مورد بررسی است. به عنوان مثال تعداد حروف فارسی ۳۲ است. بنابراین chn در محاسبه شباهت به زبان فارسی برابر ۳۲ است. $Hist_1$ و $Hist_2$ نشانگر هیستوگرام جمله مورد بررسی از متن اول و دوم می‌باشد.. $Hist_{1_i}$ برابر المان iام هیستوگرام ۱ است. همچنین $Hist_{worst}$ نیز هیستوگرام خروجی بدترین حالت ممکن است که پیش از این نحوه محاسبه آن ذکر گردید.

قابل ذکر است که ویژگی دگردیسی کلی در مقایسه دو ورودی بزرگ بسیار کارا خواهد بود. به این دلیل که این روش دارای پیچیدگی زمانی کمی است و به سرعت اجرا می‌گردد. برای دو ورودی با تعداد جملات n و m و میانگین تعداد ۱۲۰ کاراکتر در هر جمله پیچیدگی روش فوق برابر $O(mn)$ خواهد بود که نسبت

به روش‌های محاسبه شباهت مانند n-گرام و روش‌های اثر انگشت، بسیار سریع‌تر عمل می‌کند. البته این روش در کنار روش‌های مذکور می‌تواند به افزایش کارایی بیانجامد. به این صورت که با مشخص شدن یک مقدار آستانه[۱] برای تعیین کردن شباهت دو جمله، برخی از جملات که دارای شباهت بسیار زیاد هستند از متون مورد بررسی کنار گذاشته می‌شوند که خود موجب افزایش کارایی سامانه خواهد شد.

در این معیار تمامی جملات پیمایش می‌شوند و برخی از جفت جملات که شباهت آنها بیش از حد آستانه است به عنوان جفت جملات کاندید مشابه انتخاب می‌شوند. در نهایت ویژگی مشابهت جملات بر اساس معیار دگردیسی کلی به صورت زیر تعریف می‌گردد. این معیار را GDS [۲] می‌نامیم.

$$GDS\ Similarity(D_1, D_2) = \frac{S_{D_1} \cap S_{D_2}}{\min(S_{D_1}, S_{D_2})} \quad (۳-۲)$$

که در آن S_{D_1} و S_{D_2} مجموعه جملات سند اول و دوم است. در این معیار به دلیل آنکه مد نظر ما است که بخش‌های کپی شده تاثیر بیشتری روی شباهت بگذارند، در رابطه فوق از معیار Overlap استفاده نموده‌ایم.

۳-۴-۳- دگردیسی محلی[۳]

ممکن است در یک کلمه برخی از حروف جابه‌جا شده باشند و یا با یک غلط تایپی و نگارشی کوچک دیگر کلماتی که ظاهر یکسان دارند هم‌ارز قرار نگیرند. دگردیسی محلی به این معنا است که در توکن‌ها به دنبال بخش‌های جزئی‌تر مانند جایگاه حروف باشیم که ممکن است در کلمه جابه‌جا شده‌اند. برای محاسبه دگردیسی‌های محلی، یک فاصله تعریف می‌کنیم که این مقدار از ۱ تا ۵ قابل تغییر می‌باشد. این فاصله در حقیقت تعیین می‌کند کاراکتر iام یک توکن با چند کاراکتر قبل و بعد خود مقایسه شود. به عنوان مثال

---

[۱] **Threshold**

[۲] **Global deformations based sentences similarity**

[۳] **Local Deformation**

اگر مقدار فاصله برابر ۳ درنظر گرفته شود، کاراکتر iام عبارت اول با سه کاراکتر قبل و بعد عبارت دوم مقایسه می‌شود. در حین این جستجو ما ضریب تاثیر فاصله‌ای را در نظر می‌گیریم که نمایانگر آن است که هر چقدر فاصله کاراکتر iام عبارت اول، با کاراکتر $i + 3$ در عبارت دوم برابر بود، به میزان فاصله $index$ که دو کاراکتر از یکدیگر دارند مقدار تاثیر آن کم می‌شود. به زبان ساده به عنوان مثال اگر کاراکتر پنجم عبارت اول با کاراکتر پنجم عبارت دوم برابر باشد ما مقدار صفر را برای فاصله در نظر می‌گیریم و اگر کاراکتر پنجم با کاراکتر دوم از عبارت دوم برابر بود ما مقداری بیش از صفر را برای فاصله در نظر خواهیم گرفت.

در نهایت با پیمایش تمام رشته ورودی و یا زیر رشته‌های ورودی عبارات، این مقادیر فاصله با هم جمع می‌شوند. در آخر نیز این فاصله جهت نرمال سازی به ماکزیمم طول رشته اول و دوم تقسیم می‌شود.

۳-۴-۴- استخراج اثر انگشت جملات

در فصل گذشته مروری بر روش‌های مبتنی بر اثر انگشت و مدل‌های هش‌گذاری داشتیم. این الگوریتم‌ها جزو مرسوم‌ترین روش‌های کشف سرقت ادبی در زمینه تحقیقات سال‌های گذشته محسوب می‌شوند [۲۲]. در این بخش ما از تکنیک Winnowing و روش انتخاب هش n-گرام مبتنی بر کلمات (WNG) استفاده می‌نماییم.

۱-۴-۴-۳- تغییرات در الگوریتم Winnowing

در این بخش به ارائه جزئیات پیاده‌سازی الگوریتم Winnowing در محاسبه شباهت می‌پردازیم. طبق تحقیقات صورت گرفته، واحد اصلی کپی برداری در متون متوسط و بزرگ جملات هستند. در این مرحله الگوریتم Winnowing به گونه‌ای تغییر داده می‌شود که بتواند جملات کاندید دارای شباهت بسیار زیاد را نیز شناسایی نماید. مراحل اعمال روش Winnowing به صورت زیر می‌باشد.

۱. دریافت جملات جداسازی‌شده از مرحله پیش‌پردازش.

۲. ایجاد k-گرام از عبارات درون هر جمله و تبدیل آنها به هش‌ها.

۳. نمونه‌برداری از هش‌های تولید شده. نرخ نمونه‌برداری از هر مجموعه جملات متغیر و وابسته به تعداد کلمات درون مجموعه جملات است و تعداد نمونه‌های برداشت شده بین ۳۰ تا ۵۰ درصد مجموعه اصلی را پوشش می‌دهد.

۴. ذخیره‌سازی نمونه‌های منتخب از هر مجموعه به همراه اندیس گرام جملات تفکیک شده به عنوان اثر انگشت.

۵. مقایسه اثر انگشت‌های ذخیره شده از دو متن با یکدیگر.

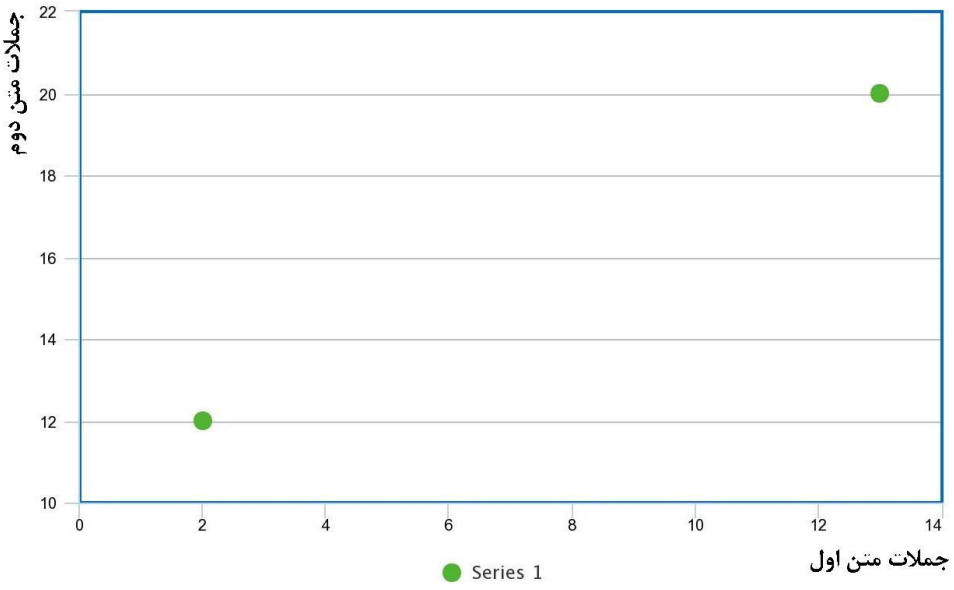
۲-۴-۴-۳- مقایسه اثر انگشت‌های دو متن

پس از مقایسه اثر انگشت‌های دو متن با سه فرضیه گوناگون مواجه می‌شویم.

در وضعیت (۱) دو متن ممکن است اثرانگشت‌های مشترکی نیز داشته باشند. اما این اثر انگشت‌های مشترک به صورت پراکنده در تمامی جملات پخش شده‌اند. شکل ۳-۳ نمونه‌ای از پراکندگی اثرانگشت‌های مشترک در دو متن با تعداد ۳۵ و ۱۴ جمله را نمایش می‌دهد. در این حالت بر حسب تصادف ممکن است که برخی از n-گرام‌ها شامل عبارات پرکاربرد فارسی شوند و برخی از اثر انگشت‌ها با یکدیگر مشابه در نظر گرفته شوند.

جدول ۳-۲ فرضیه‌های موجود در مقایسه اثر انگشت‌های مشترک دو متن

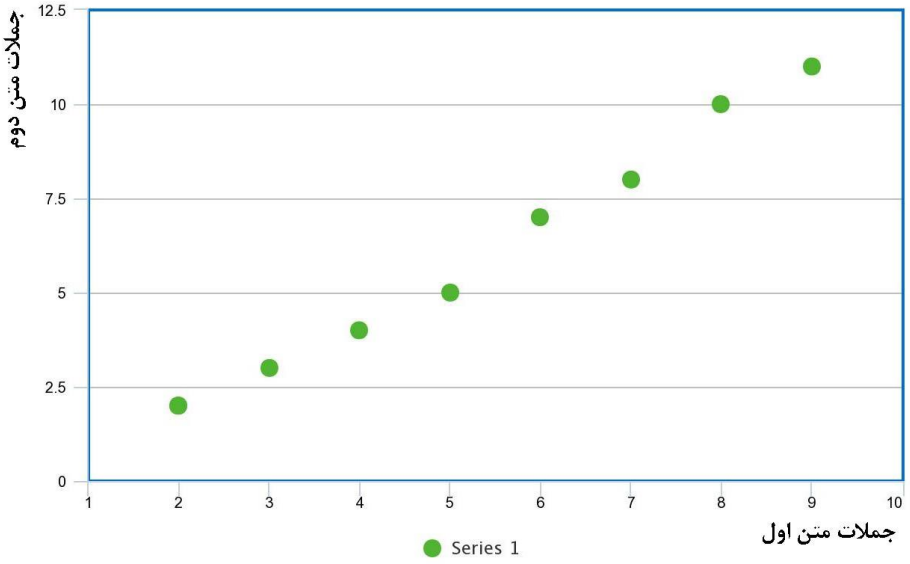
شماره	وضعیت	استنتاج
۱	دو متن اثر انگشت مشترکی ندارند و یا اثر انگشت‌های مشترک بسیار معدود هستند و به صفر میل می‌کند.	دو متن با یکدیگر هیچ شباهتی ندارند.
۲	دو متن اثر انگشت‌های مشترک بسیار زیادی دارند.	دو متن مشابه یکدیگر هستند.
۳	دو متن دارای اثر انگشت‌های نسبتاً محدود اما همگی در یک مجموعه جمله هستند.	دو متن کاملا مشابه نیستند. تنها بعضی از جملات دو متن از یکدیگر کپی‌برداری شده است.



شکل ۳-۳ پراکندگی اثرانگشت‌ها در جملات متونی که با یکدیگر شباهت ندارند

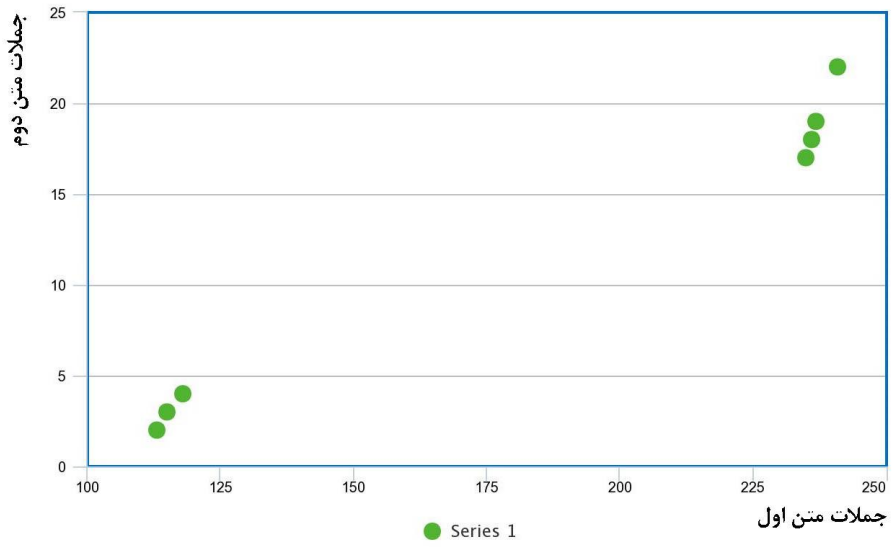
در وضعیت (۲) تقریبا تمام جملات با یکدیگر اثرانگشت‌های مشترک دارند. چرا که دو متن باید کاملا شبیه به یکدیگر باشند و حتی جملات آنها نیز با یکدیگر جابجا نشده است. بنابراین در این حالت به طور معمول تعداد جملات دو متن باهم برابر هستند. شکل ۳-۴ نمونه‌ای از این پراکندگی را نمایش می‌دهد. همانگونه که مشاهده می‌کنید تعداد جملات دو متن ۱۳ و ۱۰ عدد است. محور افق نماینده متن کپی شده است.

همانطور که مشاهده می‌کنید تمامی جملات تقریبا حتی با حفظ جایگاه کپی‌برداری شده است. تنها دو جمله کپی نشده که باعث شده نمودار از حالت خطی خارج گردد.



شکل ۳-۴ پراکندگی اثر انگشت‌ها در جملات متون کاملا متشابه

وضعیت (۳) معمولا متونی اتفاق می‌افتد که دو متن دارای اندازه جملات یکسانی نیستند. در این موارد تنها بخشی از جملات متن مبدا در متن کپی شده مشاهده می‌گردد. برای مثال در یک پژوهش ممکن است پژوهشگر برای نگارش مقاله خود از یک کتاب جملاتی را اقتباس کرده باشد. این نوع کپی برداری مرسوم‌ترین نوع کپی برداری در بین مقالات علمی-ادبی به شمار می‌رود. در شکل ۳-۵ نمونه‌ای از این موضوع نمایش داده شده است. در این شکل دو متن با اندازه جملات ۲۵ و ۲۵۰ قرار گرفته‌اند که تعداد ۷ جمله از متن بزرگ‌تر در متن کوچک‌تر کپی برداری شده است.



شکل ۳-۵ پراکندگی اثر انگشت‌ها در جملات متون نسبتاً متشابه

۳-۴-۴-۳- شباهت بین اثر انگشت‌ها و استخراج جملات کاندید مشابه

در این مرحله با استفاده از هش‌های استخراج شده و اندیس جملات همراه هر یک از این هش‌ها، جملات کاندید که بین دو متن مشترک هستند. جدول ۳-۴ چگونگی انتخاب هش‌های کاندید را از بین تمامی هش‌های ایجاد شده در جمله نمایش می‌دهد. لازم به ذکر است که ما در اینجا اندازه پنجره هش‌ها را ۴ در نظر گرفته‌ایم و هر پنجره هر بار یکی به جلو حرکت می‌کند. در هر پنجره، کوچکترین مقدار انتخاب می‌شود. هر ویژگی تنها یک بار می‌تواند به عنوان مینیمم انتخاب شود. بنابراین اگر عددی هم در پنجره خود و هم در پنجره‌های مجاور مینیمم باشد، تنها در پنجره خود انتخاب شده و پنجره‌های مجاور دیگر، مقدار مینیمم نخواهند داشت. همین مسئله موجب شده تعداد اثر انگشت‌های انتخاب شده کوچکتر از تعداد پنجره‌ها باشد. اعداد ضخیم شده که زیر آنها خط کشیده شده است همان اثر انگشت‌های منتخب هستند. در اینجا به همراه خود اثر انگشت منتخب، جایگاه جمله‌ای که کلمه در آن قرار دارد و همچنین جایگاه خود کلمه

در کل متن نیز ذخیره می‌گردد. از این دو جایگاه، به جملات منتخبی دست می‌یابیم که احتمال شباهت آنها در دو متن بیشتر از سایر بخش‌ها است.

جدول ۳-۳ مراحل انتخاب اثر انگشت‌های یک جمله و ذخیره جایگاه آنها در جمله و در کل متن

شماره	توضیحات گام هش‌نگاری	نمونه واقعی هش‌نگاری	
گام ۱	خروجی اعمال هش روی گرام کلمات یک جمله نمونه	23 45 98 13 8 88 67 42 36 71 20	
گام ۲	پنجره‌های هش با اندازه ۴ در هر جمله	(45, 98, 13, **8**)	(23, 45, 98, **13**)
		(13, 8, 88, 67)	(98, 13, 8, 88)
		(88, 67, 42, **36**)	(8, 88, 67, 42)
		(42, 36, 71, **20**)	(67, 42, 36, 71)
گام ۳	اثر انگشت‌های انتخاب شده توسط winnowing	13 8 36 20	
گام ۴	ثبت اثر انگشت به همراه موقعیت جمله و موقعیت در کل متن.	[8, $i_{statement}$, $i_{document}$] [13, $i_{statement}$, $i_{document}$] [20, $i_{statement}$, $i_{document}$] [36, $i_{statement}$, $i_{document}$]	

در انتها دو امر مهم محاسبه درصد شباهت بین دو متن و استخراج جملات و بخش‌های مشابه بین دو متن باید حصول یابند. برای محاسبه شباهت بین دو متن، از معیار جاکارد و Overlap برای تطبیق دادن شباهت بین اثر انگشت‌ها استفاده می‌کنیم که توضیحات این دو معیار در فصل ۲ آورده اشده است. ما در این پژوهش از معیار Overlap استفاده می‌نماییم. دلیل استفاده از معیار Overlap این است که این معیار معمولا قسمت‌های غیر مشترک و مازاد دو متن را کنار گذاشته و تنها بخش‌های مشترک را در نظر می‌گیرد. نمونه استفاده از این معیار در سیستم‌های پرسش و پاسخ است که در فصل گذشته مروری بر آن داشتیم [۴]. دلیل دیگر استفاده از این معیار این است که ممکن است ورودی‌های فاصله بسیار زیادی از یکدیگر داشته

باشند. برای ورودی‌های با طول رشته متفاوت نیز روش Overlap جواب معقول‌تری به ما ارائه خواهد کرد. جدول ۳-۵ نمونه واقعی از سرقت را با دو معیار متفاوت جاکارد و Overlap نمایش می‌دهد. این مثال از نمونه کپی‌برداری تکالیف درس شبکه‌های کامپیوتری استخراج شده است.

جدول ۳-۴ یک مثال دنیای واقعی در مقایسه دو معیار جاکارد و Overlap برای محاسبه شباهت اثر انگشت‌های بین دو سند

تعداد هش‌های سند D_1	۷۰۰
تعداد هش‌های سند D_2	۱۲۰
تعداد هش‌های مشترک دو سند	۷۰
معیار جاکارد	۰/۰۹
معیار Overlap	۰/۵۹

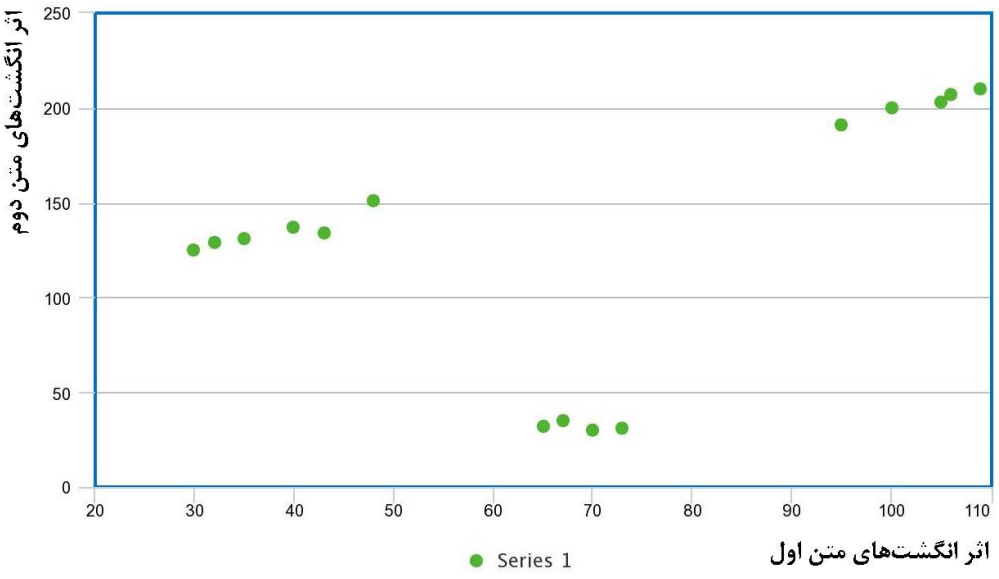
۳-۴-۴-۴- استخراج جملات و بخش‌های کاندید مشابه بین دو متن

برای استخراج جملات و بخش‌های مهم بین دو متن، دو راهکار متفاوت وجود دارد. راهکار اول بر اساس شباهت بین دو جمله استوار است. این راهکار به این صورت بیان می‌شود که اگر تعداد هش‌های مشترک بین دو جمله بیشتر از حد آستانه t باشد، آن دو جمله باهم مشابه هستند. اثبات می‌شود که پیچیدگی الگوریتم این روش با فرض تعداد اثر انگشت‌های m و n برای دو متن، برابر $O(mn)$ می‌باشد. پس از طی این مراحل جفت جملات منتخب برای مشابهت به همراه مکان آنها در بین دو متن استخراج می‌گردند. همچنین معیار شباهت مبتنی بر جملات مشترک بین دو متن به صورت زیر تعریف می‌شود.

$$Sentences\ Similarity(D_1, D_2) = \frac{S_{D_1} \cap S_{D_2}}{\min(S_{D_1}, S_{D_2})} \quad (۳-۳)$$

در رابطه فوق S_{D_1} و S_{D_2} مجموعه جملات متون اول و دوم هستند.

راهکار دیگر برای استخراج بخش‌های مشابه از دو متن استفاده از روش‌های خوشه‌بندی است [۳۸]. در این روش شباهت بین اثرانگشت‌ها را می‌توان به فضای دو بعدی که محور x و y به ترتیب اثر انگشت‌های دو متن را نمایش می‌دهد نگاشت کرد. هر نقطه در این فضا به معنی مشابهت دو اثر انگشت بین دو سند می‌باشد. به عنوان مثال در شکل ۳-۶ نمونه ای از این فضا نمایش داده شده است. تصویر تطابق اثر انگشت‌های دو متن ورودی دارای ۱۲۰ و ۲۵۰ اثر انگشت را نشان می‌دهد. نقاط سبز رنگ نماد تطابق دو اثر انگشت در دو متن می‌باشد. همانگونه که در تصویر نمایان است در سه قسمت اثر انگشت‌های مشابه سه دسته را تشکیل داده‌اند. این سه دسته نماد بخش‌هایی از متون هستند که از روی یکدیگر کپی‌برداری شده‌اند. البته ممکن است که داده‌های نویزی مانند اصطلاحات پرکاربرد زبان فارسی و یا کلیدواژه مشترک بین دو متن که به صورت مداوم در متن در حال تکرار است. این داده‌های نویزی در خوشه‌بندی در نظر گرفته نمی‌شوند.



شکل ۳-۶ فضای مشابهت اثر انگشت‌های دو متن

با استفاده از الگوریتم‌های خوشه بندی همچون الگوریتم k-میانگین[1] می‌توان این فضا را خوشه‌بندی نمود. تنها نکته‌ای که مورد نظر است این است که ما اطلاعی از تعداد خوشه‌ها نداریم. البته با بررسی نتایج و آزمایشات می‌توان نتیجه‌گیری کرد که فاصله این خوشه‌ها به چه میزان می‌تواند باشد. راهکار دوم برخی از فقدان‌های روش مبتنی بر جمله را پوشش خواهد داد. در کپی برداری و سرقت علمی-ادبی، سارق ممکن است یک جمله را به دو یا سه جمله بشکاند و یا اینکه چند جمله را به یک جمله تبدیل می‌کند.

۵-۴-۳- اختلاف تعداد جمله‌ها و واژه‌ها

این ویژگی از این جهت در نظر گرفته می‌شود که به طور معمول متونی که به طور کامل با هم شباهت دارند، دارای تعداد جمله‌ها و واژه‌های یکسانی هستند. این ویژگی همیشه نمی‌تواند ملاک ارزیابی قرار گیرد. به این صورت که اگر اختلاف واژگان و جملات دو متن زیاد باشد نمی‌توان نتیجه‌گیری کرد که دو متن شباهت ندارند، اما می‌توان اینگونه نتیجه‌گیری نمود که در صورت یکسان بودن تعداد واژگان و جملات، دو متن با احتمال بیشتری به یکدیگر شبیه هستند. از آنجا که در بخش استنتاج فازی نظر بر آن است که ویژگی‌ها نرمالسازی شده باشند، برای نرمالسازی این ویژگی آن را به صورت زیر تعریف می‌کنیم.

$$Sentences\ difference(D_1, D_2) = \frac{\min(S_{D_1}, S_{D_2})}{\max(S_{D_1}, S_{D_2})} \quad (۳-۴)$$

که در آن S_{D_1} و S_{D_2} به ترتیب تعداد جملات سند اول و سند دوم هستند.

$$Words\ difference(D_1, D_2) = \frac{\min(W_{D_1}, W_{D_2})}{\max(W_{D_1}, W_{D_2})} \quad (۳-۵)$$

که در آن W_{D_1} و W_{D_2} تعداد کلمات سند اول و سند دوم هستند.

[1] **k-means**

۳-۴-۶- استنتاج بر اساس ابعاد ورودی

در فصل پیشین با بررسی برخی از معیارها به این موضوع اشاره داشتیم که بر اساس اندازه و ابعاد رشته ورودی باید تغییری در استفاده از معیارها اعمال نماییم. برای حل این موضوع ابتدا باید مسئله را فرموله نماییم. حقایق زیر به صورت تجربی و علمی در محاسبه شباهت بین رشته‌ها ثابت شده‌اند:

۱. در محاسبه شباهت ورودی‌های کوتاه (بین ۱ تا ۲۰ کلمه) معیارهای شباهت مبتنی بر کاراکتر پاسخ بهتر و مناسب‌تری را خواهند داشت. برای محاسبه شباهت بین عبارات و ورودی‌های کوتاه الگوریتم‌های n-گرام، جارو-وینکلر و دگردیسی محلی پاسخ نزدیکتری به اعداد تخمین زده شده توسط متخصصین را به همراه دارند.

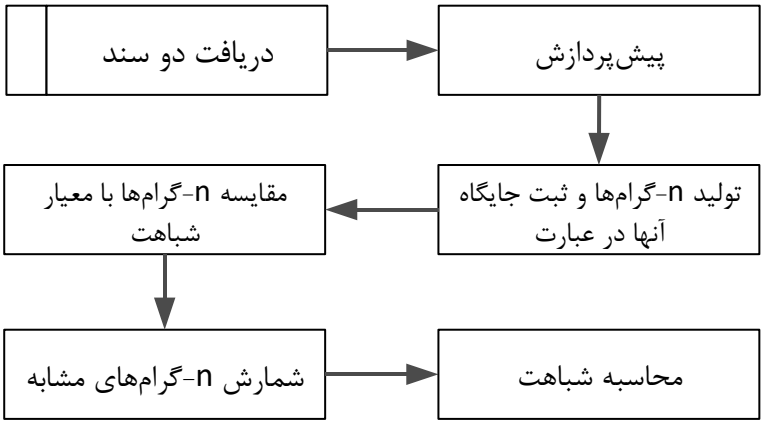
۲. در مواجهه با ورودی‌های کوچک ابتدا باید آنها را به توکن‌ها تبدیل نمود. این توکن‌ها می‌توانند مبتنی بر کاراکتر و یا کلمه باشند. سپس با مقایسه توکن‌های مشترک توسط الگوریتم‌های جاکارد و Overlap می‌توان شباهت بین آنها را محاسبه نمود. الگوریتم‌های Overlap و جاکارد در مقیاس‌های گوناگونی پاسخ متفاوتی خواهند داد.

۳. در استفاده از n-gram مبتنی بر کاراکتر، طبق تجربیات و مطالعات پیشین در مقالات گذشته، از ۴ تا ۱۴ گرام استفاده شده است. به علت کوتاه بودن عبارات، محدودیتی از نظر پردازش وجود نخواهد داشت.

۴. در سرقت علمی-ادبی در اسناد بزرگ به طور معمول قسمتی از متن حدود چندین جمله از یک متن کپی برداری می‌شود. بنابراین در اسناد بزرگ معیار شباهت مبتنی بر جمله خواهد بود.

۳-۵- محاسبه شباهت اسناد کوچک

در بخش ۳-۴-۱ اسناد را به سه دسته کوچک، متوسط و بزرگ دسته‌بندی نمودیم. در محاسبه شباهت بین اسناد کوچک نیاز است راهکارهای متفاوتی را پیش گیریم. در مقایسه این اسناد با یکدیگر نیازی به الگوریتم‌های مبتنی بر هش نیست. چرا که دلیل استفاده از الگوریتم‌های مبتنی بر هش، تبدیل فضای رشته به فضای عددی است. شکل ۳-۷ معماری این سیستم را نمایش می‌دهد.



شکل ۳-۷ معماری سیستم محاسبه شباهت اسناد کوچک

در مقیاس کوچک ورودی‌ها، به علت وجود فضای کافی برای پردازش فرآیند مشابهت بین n-گرام‌ها را بر مبنای الگوریتم‌های مشابهت رشته‌ها مانند جارو-وینکلر، LCS و دگردیسی محلی انجام می‌دهیم.

در فاز پیش‌پردازش تنها بخش حذف واژگان عمومی و ایست‌واژه‌ها را اعمال نمی‌کنیم. چرا که بر اساس تحقیقات انجام شده و نظر متخصصین، وجود واژگان عمومی و ایست‌واژه‌ها در عبارات کوتاه خود می‌تواند در خروجی شباهت تاثیرگذار باشد. بر این اساس پیش‌پردازش با طی مراحل مذکور در بخش پیشین صورت می‌گیرد. پس از فاز پیش‌پردازش، از توکن‌های ایجاد شده n-گرام‌های با ابعاد ۲، ۳ و ۴ استخراج می‌گردد. علاوه بر ثبت n-گرام‌ها جایگاه آنها در عبارت نیز به همراه آنها ذخیره خواهد شد. این جایگاه به گرام‌هایی

که در جایگاه نزدیک به یکدیگر قرار دارند وزن بیشتری می‌دهد. نکته قابل توجه این است که برای محاسبه n-گرام‌ها، از یک معیار محاسبه شباهت بین رشته استفاده می‌کنیم. این معیارها شباهت بین رشته‌ها را به صورت فازی و با عددی بین ۰ و ۱ نمایش می‌دهند. با احتساب یک حد آستانه برای استخراج گرام‌های مشابه و همچنین دخیل کردن جایگاه آنها در دو عبارت، به هر کدام از گرام‌های مشابه یک وزن داده می‌شود. بنابراین دو پارامتر وجود تکه‌های مشترک بین عبارات و جایگاه این تکه‌ها می‌توان به معیار خوبی برای محاسبه شباهت بین متون کوتاه حصول یافت.

۳-۶- سیستم استنتاج فازی برای محاسبه شباهت اسناد بزرگ

در بخش پیشین مروری بر سیستم‌های استنتاج فازی و نحوه استنتاج شباهت بر اساس این نوع سیستم‌ها داشتیم. در این بخش معماری سیستم استنتاج فازی ارائه خواهد شد. با در اختیار داشتن ویژگی‌های مورد نظر برای محاسبه شباهت بین دو متن، در این قسمت در ابتدا بایستی هر یک از این ویژگی‌ها به صورت یک مجموعه فازی نمایش داده شود. در ابتدا نیاز است متغیرهای زبانی برای نمایش میزان شباهت تعریف گردد.

۱-۶-۳- تعریف متغیرهای زبانی

در این بخش تمامی ویژگی‌ها را باید به صورت متغیرهای فازی تعریف نماییم. این متغیرهای زبانی در جدول ۳-۶ نمایش داده شده است.

جدول ۳-۵ متغیرهای زبانی تعریف شده برای نمایش مشابهت مبتنی بر دگردیسی‌های کلی جملات

متغیر زبانی	بازه عددی
متفاوت (different)	(-∞, ۰, ۰/۲۰)
نسبتاً مشابه (semi similar)	(۰, ۰/۲۰, ۰/۳۵, ۰/۵۰)
مشابه (similar)	(۰/۳۵, ۰/۵۰, ∞, ∞)

جدول ۳-۶ متغیرهای زبانی تعریف شده برای نمایش مشابهت بین اثر انگشت‌های بین دو متن

متغیر زبانی	بازه عددی
متفاوت (different)	(-∞, -∞, ۰/۱۰, ۰/۳۰)
نسبتاً مشابه (semi similar)	(۰/۱۰, ۰/۳۰, ۰/۵۰, ۰/۶۵)
مشابه (similar)	(۰/۵۰, ۰/۶۵, ∞, ∞)

جدول ۳-۷ متغیرهای زبانی تعریف شده برای نمایش مشابهت بین جملات مبتنی بر اثر انگشت‌های مشترک جملات

متغیر زبانی	بازه عددی
متفاوت (different)	(-∞, ۰, ۰/۲۰)
نسبتاً مشابه (semi similar)	(۰, ۰/۲۰, ۰/۳۵, ۰/۵۰)
مشابه (similar)	(۰/۳۵, ۰/۵۰, ∞, ∞)

جدول ۳-۸ متغیرهای زبانی تعریف شده برای نمایش ویژگی اختلاف کلمات و جمله‌ها

متغیر زبانی	بازه عددی
اختلاف زیاد (high diff)	(-∞, -∞, ۰/۴۰, ۰/۶۰)
اختلاف متوسط (medium diff)	(۰/۴۰, ۰/۶۰, ۰/۷۰, ۰/۹۰)
اختلاف کم (low diff)	(۰/۷۰, ۰/۹۰, ∞, ∞)

جدول ۳-۹ متغیرهای زبانی تعریف شده برای نمایش میزان شباهت خروجی (تالی)

متغیر زبانی	بازه عددی
متفاوت (different)	(-∞, -∞, ۰/۱۵, ۰/۳۵)
نسبتاً مشابه (semi similar)	(۰/۱۰, ۰/۳۵, ۰/۵۰, ۰/۷۰)
مشابه (similar)	(۰/۵۰, ۰/۷۰, ∞, ∞)

جدول ۳-۱۰ متغیرهای سیستم استنتاج فازی محاسبه شباهت

نام مجموعه فازی	حرف اختصاری
شباهت جملات مبتنی بر معیار دگردیسی کلی	GDS[۱]
شباهت بین اثر انگشت‌های دو متن	FP[۲]
شباهت جملات مبتنی بر معیار اثر انگشت‌های مشترک جمله	FPS[۳]
اختلاف تعداد کلمات	SD[۴]
اختلاف تعداد جملات	WD[۵]

[۱] **Global Deformations based Sentences Similarity**

[۲] **Finger print**

[۳] **Finger print sentences similarity**

[۴] **Sentences Difference**

[۵] **Words Difference**

۲-۶-۳- پایگاه قواعد فازی

پس از تعریف متغیرهای فازی، نیاز به مجموعه قواعدی داریم که بتواند غالب حالات سیستم استنتاج فازی را پوشش دهد. مجموعه قواعد سیستم استنتاج فازی این سیستم با استفاده از بررسی‌های انجام شده در سرقت‌های پیکره PAN و پیکره ارزیابی شباهت بین عبارات صورت گرفته است. همچنین برای تطبیق سندیت قواعد با مثال‌های دنیای واقعی از مثال واقعی پاسخ تکالیف درسی دانشکده کامپیوتر دانشگاه صنعتی شاهرود و شباهت بین آنها نیز استفاده شده است. این مجموعه قواعد در جدول ۳-۱۲ گرد آورده شده است.

جدول ۳-۱۱ قواعد فازی تعریف شده در سیستم استنتاج محاسبه شباهت دو سند

شماره قاعده	نام متغیر					نتیجه سیستم
	GDS	FP	FPS	SD	WD	
۱	مشابه	مشابه	مشابه	کم	کم	مشابه
۲	مشابه	مشابه	مشابه	کم	زیاد	مشابه
۳	مشابه	مشابه	مشابه	زیاد	کم	مشابه
۴	متفاوت	متفاوت	متفاوت	کم	کم	متفاوت
۵	متفاوت	نسبتاً مشابه	متفاوت	زیاد	زیاد	متفاوت
۶	متفاوت	مشابه	متفاوت	زیاد	زیاد	نسبتاً مشابه
۷	مشابه	-	-	کم	-	مشابه
۸	متفاوت	-	متفاوت	زیاد	-	متفاوت
۹	متفاوت	متفاوت	-	-	زیاد	متفاوت
۱۰	متفاوت	متفاوت	-	-	-	متفاوت
۱۱	مشابه	متفاوت	-	-	-	نسبتاً مشابه
۱۲	متفاوت	مشابه	-	-	-	نسبتاً مشابه
۱۳	مشابه	نسبتاً مشابه	-	-	-	مشابه
۱۴	نسبتاً مشابه	مشابه	-	-	-	مشابه

۱۵	نسبتاً مشابه	متفاوت	–	–	–	نسبتاً مشابه
۱۶	متفاوت	نسبتاً مشابه	–	–	–	نسبتاً مشابه
۱۷	نسبتاً مشابه	نسبتاً مشابه	–	–	–	نسبتاً مشابه
۱۸	مشابه	مشابه	–	–	–	مشابه
۱۹	–	متفاوت	متفاوت	–	–	متفاوت
۲۰	–	مشابه	متفاوت	–	–	نسبتاً مشابه
۲۱		متفاوت	مشابه	–	–	نسبتاً مشابه
۲۲	–	مشابه	نسبتاً مشابه	–	–	مشابه
۲۳	–	نسبتاً مشابه	مشابه	–	–	مشابه
۲۴	–	نسبتاً مشابه	متفاوت	–	–	نسبتاً مشابه
۲۵	–	متفاوت	نسبتاً مشابه	–	–	نسبتاً مشابه
۲۶	–	نسبتاً مشابه	نسبتاً مشابه	–	–	نسبتاً مشابه
۲۷	–	مشابه	مشابه	–	–	مشابه
۲۸	متفاوت	–	متفاوت	–	–	متفاوت
۲۹	مشابه	–	متفاوت	–	–	نسبتاً مشابه
۳۰	متفاوت	–	مشابه	–	–	نسبتاً مشابه
۳۱	مشابه	–	نسبتاً مشابه	–	–	نسبتاً مشابه
۳۲	نسبتاً مشابه	–	مشابه	–	–	مشابه
۳۳	نسبتاً مشابه	–	متفاوت	–	–	نسبتاً مشابه
۳۴	متفاوت	–	نسبتاً مشابه	–	–	نسبتاً مشابه
۳۵	نسبتاً مشابه	–	نسبتاً مشابه	–	–	نسبتاً مشابه
۳۶	مشابه	–	مشابه	–	–	مشابه
۳۷	نسبتاً مشابه	نسبتاً مشابه	نسبتاً مشابه	–	–	نسبتاً مشابه
۳۸	نسبتاً مشابه	مشابه	نسبتاً مشابه	–	–	نسبتاً مشابه
۳۹	مشابه	نسبتاً مشابه	نسبتاً مشابه	–	–	نسبتاً مشابه

۴۰	نسبتاً مشابه	نسبتاً مشابه	مشابه	–	–	نسبتاً مشابه
۴۱	مشابه	نسبتاً مشابه	مشابه	–	–	مشابه
۴۲	مشابه	مشابه	نسبتاً مشابه	–	–	مشابه
۴۳	نسبتاً مشابه	مشابه	مشابه	–	–	مشابه
۴۴	نسبتاً مشابه	متفاوت	متفاوت	–	–	متفاوت
۴۵	متفاوت	نسبتاً مشابه	متفاوت	–	–	متفاوت
۴۶	متفاوت	متفاوت	نسبتاً مشابه	–	–	متفاوت
۴۷	متفاوت	نسبتاً مشابه	نسبتاً مشابه	–	–	نسبتاً مشابه
۴۸	نسبتاً مشابه	متفاوت	نسبتاً مشابه	–	–	نسبتاً مشابه
۴۹	نسبتاً مشابه	نسبتاً مشابه	متفاوت	–	–	نسبتاً مشابه

این نکته حائز اهمیت است که تمامی حالات و قواعد مشتمل بر سیستم استنتاج فازی فوق براب 3^5 حالت خواهد بود. اما دو متغیر اختلاف جملات و کلمات می‌توانند از قسمت اعظمی از این فضا حذف گردند. همچنین لازم به ذکر است که در جدول فوق تمامی حالات برای سه ویژگی دگردیسی کلی (GDS)، اثر انگشت‌های متون (FP) و جملات مشترک مبتنی بر اثر انگشت‌های متون (FPS) در نظر گرفته شده است.

پس از تعریف قواعد استنتاج در سیستم فازی فوق، این سیستم به صورت واقعی برای آزمایشات و بررسی نتایج در محیط متلب و جعبه‌ابزار طراحی منطق فازی[۱] پیاده‌سازی گردید. پس از انجام آزمایشات، تمامی این بخش‌ها با استفاده از کتابخانه Accord-Net در زبان سی‌شارپ پیاده‌سازی شده است. در نهایت با اعمال دو ورودی متوسط یا بزرگ ویژگی‌های اصلی از آنها استخراج شده و طی اعمال سیستم استنتاج فازی، خروجی به صورت عددی بین صفر و یک نمایش داده خواهد شد. این عدد نشانگر شباهت بین دو ورودی

[۱] **Fuzzy logic designer toolbox**

است. با محاسبه درجه عضویت این عدد در تالی می‌توان دریافت که این دو متن ورودی در کدامیک از سه کلاس مشابه، نسبتاً مشابه و متفاوت قرار دارند.

۳-۷- استخراج جملات کاندید مشابه دو سند

پس از محاسبه شباهت بین دو سند توسط سیستم استنتاج فازی تشریح شده، برای استخراج و نمایش جملات کاندید مشابه بین دو سند توسط دو معیار محاسبه دگردیسی‌های کلی سطح جمله و معیار شباهت جملات مبتنی بر اثر انگشت‌های مشترک بین جملات، مراحل زیر را پیش‌رو خواهیم داشت:

۱. دریافت جملات تفکیک شده و استخراج اثرانگشت‌های n-گرام‌ها و هیستوگرام کاراکترهای آنها.

۲. بررسی تشابه بین جملات با اعمال معیار دگردیسی کلی: در این معیار در ابتدا جفت جملات با یکدیگر مقایسه می‌شوند و برای هر جفت جمله معیار دگردیسی محاسبه می‌شود. دو راهکار برای استخراج جملات کاندید وجود دارد. راهکار اول به این صورت خواهد بود که جفت جملات بر اساس شباهت بین آنها مرتب می‌شوند و تعداد P جفت جمله با بالاترین عدد شباهت استخراج می‌گردند. راهکار دوم این است که یک حد آستانه t تعریف شود و هر جفت جمله بیش از حد آستانه به عنوان جفت جملات کاندید در نظر گرفته شود. ما در این پژوهش راهکار اول را پیاده‌سازی نموده‌ایم.

۳. بررسی تشابه جفت جملات با اعمال معیار شباهت جملات مبتنی بر اثر انگشت‌های مشترک: در این معیار در ابتدا بایستی جملات دو متن یک به یک با یکدیگر مقایسه شوند. در هر جمله تعدادی اثر انگشت منتخب وجود دارد که این اثر انگشت‌ها در جملات با اثر انگشت‌های جمله مقابل مقایسه می‌شوند. در این مرحله نیز دو راهکار متفاوت وجود دارد. راهکار اول به این صورت خواهد بود که تعداد اثر انگشت‌های مشترک بین جفت جملات ذخیره گردد. در نهایت این تعداد به صورت نزولی

مرتب شده و P جمله اول به عنوان جفت جملات کاندید استخراج می‌شوند. راهکار دیگر تعریف آستانه t است به این نحو که جفت جملاتی که دارای اثر انگشت‌های مشترک بیش از حد آستانه باشند به عنوان جفت جملات کاندید انتخاب می‌گردند.

۳-۸- نتیجه‌گیری

در این بخش به معرفی معماری کلی سیستم استنتاج محاسبه شباهت بین عبارات پرداختیم. این معماری دارای سه بخش اصلی پیش‌پردازش، انتخاب ویژگی و استنتاج فازی است. در بخش پیش‌پردازش به چگونگی آماده‌سازی متون فارسی پرداختیم. برخی از موارد مانند حذف اعداد و جداسازی جملات از جمله اعمال در این بخش بود. در بخش استخراج ویژگی‌ها دو معیار دگردیسی کلی و هش‌گذاری ویژگی‌های متن مطرح گشت. در نهایت در بخش استنتاج فازی تمامی این ویژگی‌ها به صورت قواعد فازی به سیستم منتقل شد. سیستم با دریافت عبارات، شباهت را بر اساس وافازی‌گری به صورت عددی بین صفر و یک نمایش می‌دهد. در انتها روش استخراج جفت جملات کاندید شباهت بین دو متن ارائه گشت.

فصل چهارم

جزئیات پیکربندی و ارزیابی روش پیشنهادی

٤- جزئیات پیکربندی و ارزیابی روش پیشنهادی

در فصل سوم، سیستم محاسبه شباهت بین متون معرفی شد. در این فصل ابتدا به معرفی معیارهای ارزیابی مورد استفاده و پایگاه داده مورد ارزیابی در این پایان‌نامه پرداخته می‌شود. سپس نتایج حاصل از اعمال روش پیشنهادی بر روی داده‌ها ارائه می‌گردد.

۴-۱- پایگاه داده مورد ارزیابی

پایگاه‌های داده مختلفی در زمینه سرقت علمی ادبی به زبان انگلیسی منتشر شده‌اند. این پایگاه‌های داده انواع و اقسام مشابهت از مشابهت لغوی تا مشابهت‌های ساختاری و معنایی را در خود جای داده‌اند. در زبان فارسی نیز دو پایگاه داده پیکره فارسی PAN و پیکره سمیم برای ارزیابی سیستم‌های کشف سرقت ادبی منتشر شده است. در نهایت با بررسی‌های انجام شده در این پایگاه‌های داده از پیکره فارسی PAN که از اسناد موجود در ویکی‌پدیا تشکیل شده است، و همچنین پیکره ارزیابی شباهت بین عبارات که توسط آزمایشگاه وب‌کاوی دانشگاه صنعتی شاهرود توسعه داده شده است استفاده خواهیم نمود.

۴-۱-۱- پیکره تک‌زبانه فارسی ارزیابی سامانه‌های تشخیص سرقت ادبی (PAN)

مجموعه پایگاه‌های داده PAN یکی از معروف‌ترین پایگاه‌های داده‌ها به زبان انگلیسی در زمینه ارزیابی سیستم‌های کشف سرقت ادبی هستند. این پایگاه داده هر دو سال یک بار توسط پژوهشکده فناوری و

اطلاعات[1] منتشر می‌شود. در این پایگاه داده سرقت علمی تک زبانه شامل ۴ سطح مختلف می‌باشد. این ۴ بخش شامل موارد ذیل هستند.

۱. **نزدیک به کپی:** در این کلاس در سند مشابه قسمت‌هایی از متن بدون تغییر از متن اصلی آورده شده است.

۲. **نسخه‌برداری سبک:** در این کلاس سند مشابه از متن اصلی با وارد کردن تغییرات جزئی (استفاده از کلمات مترادف و تغییر دستوری) استفاده کرده است.

۳. **نسخه‌برداری سنگین:** در این کلاس سند از نسخه اصلی بازنویسی کاملی را با تغییر ساختار انجام داده است.

۴. **بدون سرقت علمی:** در این کلاس سند از نسخه اصلی در تعداد کوتاهی کلمه مانند گزاره ۲۰۰ تا ۳۰۰ کلمه‌ای استفاده کرده است [۹].

نمونه فارسی این پیکره در سال ۲۰۱۵ توسط خوشنوا و همکارانش [۳] تکمیل گشته است. این پیکره بیش از ۳۰۰۰ سند فارسی دارد. اسناد مورد استفاده در تهیه پیکره از اسناد فارسی ویکی‌پدیا انتخاب شده است. صفحات فارسی ویکی‌پدیا سال ۲۰۱۵ میلادی خزش[2] شده است و محتوای هر یک از اسناد شامل محتوای متنی صفحات ویکی‌پدیا می‌باشد.

بخشی از اسناد به عنوان اسناد مرجع یا به عبارت دیگر، اسناد پایه جهت استخراج موارد سرقت علمی مورد استفاده قرار گرفته‌اند. بخشی دیگر نیز اسناد مشکوک به تقلب می‌باشند و موارد سرقت علمی در آن‌ها درج

---

[1] **ICT Research Institute**

[2] **Crawl**

شده است. در نسخه سوم این پیکره، اسناد مرجع و مشکوک هر یک شامل ۱۵۰۰ سند می‌باشد. جزئیات این پیکره در جدول زیر نمایش داده شده است.

جدول ۴-۱ جزئیات اسناد پیکره تک‌زبانه فارسی ارزیابی سامانه‌های تشخیص سرقت ادبی

اسناد مرجع	۱۵۲۴
اسناد مشکوک	۱۵۰۰
اسناد بدون تقلب	۷۳۶
اسناد با تقلب	۲۲۶۲
کل اسناد	**۳۰۲۵**

در پیکره فوق برای تولید موارد تقلب از فرآیند ابهام‌افزایی "مصنوعی" استفاده گردیده است. در این بخش، عملیاتی چون جابجایی کلمات، حذف و اضافه نمودن کلمات و جایگزین نمودن کلمات با کلمات هم‌معنا اعمال شده است [3].

جدول ۴-۲ موارد ابهام افزایی پیکره تک‌زبانه فارسی ارزیابی سامانه‌های تشخیص سرقت ادبی

ابهام‌افزایی مصنوعی	۱۷۱۷
ابهام‌افزایی کم	۹۴۰
ابهام‌افزایی زیاد	۷۷۷
ابهام‌افزایی شبیه‌سازی شده	۱۶۷
بدون ابهام‌افزایی	۱۸۴
کل موارد سرقت علمی	۲۰۶۸
متوسط طول موارد سرقت شده	۵ جمله

نحوه ابهام‌افزایی اطلاعات به این صورت است که بخش‌های مختلفی از سند منبع با در نظر گرفتن مکان و طول رشته به سند مشکوک انتقال داده شده است. طول این رشته‌ها معمولا بین ۳۰ تا ۵۰۰ کلمه است.

فایل *XML* شامل اطلاعات سند منبع و سند مشکوک در قالب مشخص در یک سری فایل جداگانه قرار گرفته است. همچنین اطلاعات سرقت‌های به کار رفته، مکان رشته‌ها، طول رشته‌ها و نسبت سرقت نیز در این فایل موجود است. این فایل به فرمت زیر برای تمامی اسناد مشکوک ارائه شده است.

```
<document reference="suspicious-document-536.txt">
<feature
      name="plagiarism"
      obfuscation="random"
      obfuscation_degree="high"
      source_length="192"
      source_offset="46"
      source_reference="source-document00001394685.txt"
      this_length="196"
      this_offset="3863"
/>
</document>
```

قابل ذکر است اگر چندین مورد سرقت از سند منبع صورت گرفته باشد چندین تگ[۱] ***feature*** در سند فوق قرار می‌گیرد که هر یک دارای اطلاعات سرقت مربوط به خود است که در متاتگ خاص خود قرار می‌گیرد. در سرقت‌های ذکر شده در فایل *XML*، تگ ***obfuscation*** می‌تواند ۵ مقدار مختلف اختیار کند. این مقادیر به شرح زیر می‌باشند:

۱. بدون ابهام‌زایی[۲]: بخش مشخص شده از منبع بدون هیچ تغییری در سند مشکوک قرار داده می‌شود.

---

[۱] **Tag**

[۲] **No Obfuscation**

۲. تغییر جایگاه کلمات به صورت تصادفی: بخش مشخص شده از منبع با تغییر جایگاه کلمات به صورت تصادفی در سند مشکوک قرار داده می‌شود.

۳. اتصال کلمات با حفظ جایگاه POS[۱]: برخی از کلمات بخش مشخص شده از منبع با کنار گذاشتن قسمت‌های گفتاری، به یکدیگر چسبیده و در سند مشکوک قرار می‌گیرند.

۴. جایگذاری مترادف[۲] کلمات: برخی از کلمات بخش مشخص شده از منبع با مترادف خود جایگزین شده و در سند مشکوک قرار می‌گیرند.

۵. افزودن و حذف: بخش مشخص شده از منبع با افزودن و یا حذف برخی قسمت‌ها در سند مشکوک قرار می‌گیرد. افزودن و حذف کلمات به صورت تصادفی اتفاق می‌افتد.

تعداد عملات انجام شده بر روی بخش منبع درجه ابهام را تعیین می‌کند. در این پیکره درجه‌های مختلف ابهام همچون "بدون ابهام"، "کم"، "متوسط" و "زیاد" است.

۴-۱-۲- پیکره ارزیابی شباهت بین عبارات[۳]

این پیکره برای ارزیابی سیستم محاسبه شباهت فازی تولید شده است. این پیکره شامل ۷۶۰ جفت عبارت است که شباهت آنها توسط دو متغیر زبانی و عدد شباهت قابل نمایش می‌باشد. این جفت عبارات شامل جملات کوتاه، پاراگراف‌ها و متون طولانی تا سه هزار کلمه است. بخشی از این پیکره از تکالیف درسی دانشجویان دانشکده کامپیوتر دانشگاه صنعتی شاهرود اقتباس شده‌است. از آنجا که برخی از تکالیف درسی

[۱] **Part of speech**

[۲] **Synonym**

[۳] نسخه ۱: ایجاد شده در سال ۱۳۹۵، آزمایشگاه وب‌کاوی دانشکده کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود.

شامل سرقت ادبی ناشیانه در محدوده جملات بوده‌اند، از این مجموعه برای تولید پیکره استفاده شده است. این پیکره در یک فایل *XML* گردآوری شده است. تقسیم‌بندی این پیکره در جدول زیر آورده شده است.

جدول ۴-۳ جزئیات کلاس‌های موجود در پیکره ارزیابی شباهت بین عبارات

متفاوت	۲۳۰
نسبتاً مشابه	۳۵۰
اسناد بدون تقلب	۱۸۰

۴-۲- ساز و کار ارزیابی

با توجه به ساختار پایگاه داده تشریح شده در بخش پیشین، نحوه ارزیابی سیستم محاسبه شباهت در این پژوهش بر اساس این پایگاه داده مدون خواهد شد. برای ارزیابی سیستم به تمامی داده‌های منبع که شامل ۱۵۰۰ سند است نیاز خواهیم داشت. ما از مجموعه فایل‌های *XML* شامل اطلاعات اسناد مشکوک و سرقت‌های صورت گرفته در آنها برای مشخص نمودن جفت اسنادهای مورد بررسی استفاده خواهیم کرد. در بین ۱۵۰۰ فایل *XML* شامل اطلاعات اسناد مشکوک ۷۶۶ سند دارای سرقت هستند. این نکته حائز اهمیت است که در هر یک از فایل‌های *XML* شامل اطلاعات سند مشکوک ممکن است بیش از یک سرقت از منابع مختلف رخ داده باشد. در مجموع ۳۵۱ سند بدون ابهام‌افزایی، ۹۴۰ سرقت با ابهام‌افزایی کم و ۷۷۷ مورد با ابهام‌افزایی زیاد رخ داده است. در مجموع ۲۰۶۸ جفت سند به سرقت رفته وجود دارد که از آنها برای ارزیابی سیستم استفاده خواهیم کرد. بر اساس طول رشته به سرقت رفته و طول کل رشته سند مشکوک این اسناد به دو دسته مشابه و نسبتاً مشابه تقسیم می‌شوند. برای ارزیابی صحت عملکرد سیستم استنتاج فازی نیاز است پیکره PAN به سه کلاس مختلف متفاوت، نسبتاً مشابه و مشابه تقسیم‌بندی شود.

با استفاده از اطلاعات فایل XML سند مشکوک این امر قابل حصول می‌باشد. در مراحل زیر روند کلاسه‌بندی جفت اسناد مرجع و مشکوک را تشریح می‌نماییم:

۱. اسناد مرجع که در فایل XML به عنوان مبدا سرقت مشخص شده‌اند یک به یک با سند مشکوک ارزیابی می‌شوند.

۲. پس از محاسبه شباهت بین سند مرجع و مشکوک، با استفاده از متاتگ‌های موجود در فایل XML طول رشته کپی شده در مبدا و مقصد استخراج می‌گردد. از طول این دو رشته میانگین گرفته می‌شود و آن را $S_{plagiarism}$ می‌نامیم. طی پردازشی ساده طول کل رشته مبدا و مقصد استخراج شده و آنها را S_{source} و $S_{suspicious}$ می‌نامیم.

۳. با در اختیار داشتن این پارامترها، میزان کپی‌برداری در سند مشکوک را محاسبه می‌نماییم:

$$Plagiarism\ Percentage = \frac{S_{plagiarism}}{Mean(S_{suspicious}, S_{source})} \quad (۴-۱)$$

با آستانه‌گذاری خروجی عدد شباهت را به یک کلاس نگاشت خواهیم داد. این آستانه‌ها به صورت زیر تعریف می‌شوند. بر اساس ارزیابی‌های انجام شده توسط متخصصین و همچنین پارامترهای در نظر گرفته شده در متغیر فازی تالی محاسبه شباهت، این آستانه ۰.۶۵ انتخاب شده است. بدیهی است خروجی کمتر از ۰.۶۵ در کلاس نسبتاً مشابه و بیش از ۰.۶۵ در کلاس مشابه قرار می‌گیرد. به دلیل اینکه قطعا در این متون تقلبی صورت گرفته است این بخش‌ها فاقد کلاس متفاوت هستند و باقی اسناد موجود در پیکره که در آن سرقتی صورت نگرفته است در کلاس متفاوت قرار می‌گیرد.

در نهایت با دسته‌بندی تمام مجموعه‌ها تعداد جفت اسناد موجود پیکره PAN در کلاس‌های گوناگون به صورت زیر خواهد بود.

جدول ۴-۴ تعداد اسناد موجود پیکره PAN در سه کلاس متفاوت جهت ارزیابی سیستم

متفاوت	۷۳۶
نسبتاً مشابه	۱۷۳۶
مشابه	۳۳۲

۴-۳- معرفی معیارهای ارزیابی

معیارهای ارزیابی برای نشان دادن کارایی الگوریتم و نیز مقایسه روش پیشنهادی با سایر روش‌ها، به کار می‌روند. معیارهای دقت[۱]، *Precision*، *Recall* و معیار *F* پرکاربردترین معیارها در ارزیابی سیستم‌ها می‌باشند. برای معرفی معیارهای ارزیابی، نیاز به معرفی مفاهیم و پارامترهایی داریم که از آنها در معیارهای فوق استفاده می‌شود. این مفاهیم در جدول ۴-۵ آورده شده‌اند.

جدول ۴-۵ پارامترهای مورد نیاز اندازه‌گیری معیارهای کارایی

		کلاس پیش‌بینی شده	
		نسبت داده نشده به دسته C_i	نسبت داده شده به کلاس C_i
کلاس واقعی	متعل به کلاس C_i	FN_i	TP_i
	متعلق به کلاس غیر از C_i	TN_i	FP_i

[۱] **Accuracy**

در جدول فوق به علت سهولت فهم، دسته‌ها به صورت پارامتر i در نظر گرفته شده‌اند. در حقیقت ما سه کلاس مختلف مشابه، نسبتاً مشابه و متفاوت خواهیم داشت که هر یک دارای پارامترهای زیر هستند. برای نمونه پارامترهای مربوط به کلاس مشابه در جدول زیر گردآوری شده‌است. بدیهی است برای دو کلاس دیگر نیز این پارامترها به صورت مجزا تعریف می‌شوند.

جدول ۴-۶ پارامترهای مورد نیاز اندازه گیری معیار کارایی کلاس "مشابه"

پارامتر	توضیحات
TP_S	جفت متون در دسته مشابه که توسط سیستم مشابه تشخیص داده شده‌اند.
FP_S	جفت متون در دسته‌های غیر از مشابه که توسط سیستم مشابه تشخیص داده شده‌اند.
FN_S	جفت متون در دسته مشابه که توسط سیستم مشابه تشخیص داده نشده‌اند.
TP_S	جفت متون در دسته‌های غیر از مشابه که توسط سیستم مشابه تشخیص داده نشده‌اند.

پس از تعریف مفاهیم اصلی، معیارهای ارزیابی الگوریتم به شرح ذیل می‌باشد:

معیار دقت: نسبت نمونه‌های درست تشخیص داده شده را به کل نمونه‌ها نشان می‌دهد (رابطه ۴-۲):

$$Accuracy: A_i = \frac{TP_i+TN_i}{TP_i+TN_i+FP_i+FN_i} \quad (۴-۲)$$

معیار Precision: این معیار، نسبت تعداد نمونه درست نسبت داده شده را به کل نمونه‌های نسبت داده شده به همان کلاس نشان می‌دهد. مقدار معیار Precision، درجه پایداری[۱] الگوریتم مورد بررسی را نشان می‌دهد.

[۱] **Soundness**

$$Precision: P_i = \frac{TP_i}{TP_i+FP_i} \quad (۴-۳)$$

معیار *Recall*: معیار بازخوانی نسبت تعداد نمونه‌هایی را که به درستی در یک کلاس قرار گرفته‌اند به کل نمونه‌هایی که بایستی در آن کلاس قرار می‌گرفتند، نشان می‌دهد. این معیار، درجه تمامیت[۱] الگوریتم مورد بررسی را نشان می‌دهد.

$$Recall: R_i = \frac{TP_i}{TP_i+FN_i} \quad (۴-۴)$$

معیار F_β: معیاری است که از ترکیب دو معیار Precision و Recall بدست می‌آید. چرا که معیارهای صحبت و بازخوانی به تنهایی برای اندازه‌گیری کارایی سیستم محاسبه شباهت متون موثر نبوده و استفاده مجزای آنها ممکن است نتایج نادرستی از سیستم را ارائه دهد. از این رو لازم است که این معیارها به گونه‌ای با یکدیگر ترکیب شوند. این معیار میزان تاثیر معیار میزان صحبت و بازخوانی را در ارزیابی سیستم محاسبه شباهت در نظر می‌گیرد.

$$F_\beta = \frac{(\beta^2+1)*Recall*Precision}{\beta^2\,(Recall+Precision)} \quad (۴-۵)$$

در رابطه فوق معمولا مقدار β برابر ۱ در نظر گرفته می‌شود. معنای مقدار ۱ برای β به معنی تاثیری برابر از صحت و بازخوانی خواهد بود. رابطه زیر این معیار را نمایش می‌دهد.

$$F_{measure} = \frac{2*Recall*Precision}{Recall+Precision} \quad (۴-۶)$$

[۱] **Completness**

۴-۴- نتایج محاسبه شباهت اسناد کوچک

در فصل پیشین معیار محاسبه شباهت با دخالت جایگاه گرام‌ها در عبارات کوچک معرفی شد. در این بخش پارامترهای مختلفی همچون، الگوریتم محاسبه شباهت و عدد n برای تولید گرام‌ها را مورد بررسی قرار خواهیم داد. قابل ذکر است که محاسبه اسناد کوچک بر روی بخشی از پیکره ارزیابی شباهت آزمایشگاه وب‌کاوی دانشگاه صنعتی شاهرود با تعداد ۳۰۰ جفت عبارت با مقیاس کوچک ارزیابی شده است. هر یک از این جفت عبارات دارای یک عدد شباهت بین ۰ و ۱ هستند. هر چه شباهت بین دو عبارت بیشتر باشد این عدد به ۱ نزدیک‌تر است.

در ارزیابی الگوریتم‌های محاسبه شباهت بین دو گرام، از معیار فاصله اقلیدسی استفاده نموده‌ایم. این معیار با محاسبه خطای خروجی با عدد واقعی شباهت، مجذور مجموعه شباهت‌ها را به عنوان خطای خروجی نمایش می‌دهد. برای یکسان بودن فضای آزمایش برای تمامی الگوریتم‌ها، آزمایشات با در نظر گرفتن عدد n برابر ۳ برای n-گرام‌ها و حد آستانه ۰,۸ برای تمامی الگوریتم‌ها صورت گرفته‌است. جدول زیر این مقدار خطا را برای الگوریتم‌های مورد بررسی محاسبه نمایش می‌دهد.

جدول ۴-۷ مجموعه خطای تفاضل بین عدد محاسبه شده با فاصله واقعی شباهت با استفاده از معیار فاصله اقلیدسی

معیار مورد نظر	خطای خروجی
فاصله اقلیدسی (نرمال شده)	۱۱۶/۸
LCS	۵۰/۱
Levenshtein	۲۷/۹
Jaro-Winkler	۲۷/۸
Local-Deformations (2)	۲۹/۵

Local-Deformations (4)	۲۶/۳

بر اساس موارد نتیجه‌گیری شده فوق ما از معیار جارو-وینکلر برای محاسبه شباهت بین گرام‌ها استفاده نموده‌ایم. هرچند که معیار دگردیسی محلی نیز می‌تواند معیار موثری باشد، اما این معیار در همسایگی‌های بیش از دو ممکن است در محاسبه شباهت بین دو عبارت نامشابه با خطا همراه باشد.

جدول ۴-۸ مجموعه خطای بین عدد محاسبه شده با فاصله واقعی شباهت با استفاده از معیار فاصله اقلیدسی

طول گرام‌ها	خطای خروجی
۲	۱۶/۷
۳	۲۳/۸
۴	۳۱/۱

۴-۵- نتایج سیستم استنتاج فازی در محاسبه شباهت

در این بخش به نتایج عددی روش پیشنهادی با داده‌های دو پیکره مورد بررسی می‌پردازیم. جدول ۴-۹ نتایج ماتریس درهم‌ریختگی[۱] روش پیشنهادی را برای دو پیکره مورد بررسی نشان می‌دهد. برای محاسبه معیارهای کارایی کل، میانگین کارایی هر یک از کلاس‌های زیر محاسبه شده است.

جدول ۴-۹ ارزیابی نتایج محاسبه شباهت بر روی پیکره PAN

		کلاس پیش‌بینی شده		
		مشابه	نسبتاً مشابه	متفاوت
کلاس واقعی	مشابه	۲۴۷	۴۱	۴۴
	نسبتاً مشابه	۱۰۸	۹۶۱	۶۶۷

[۱] **Confusion matrix**

	متفاوت	۰	۰	۷۳۶

جدول ۴-۱۰ ارزیابی نتایج محاسبه شباهت بر روی پیکره ارزیابی شباهت بین عبارات

		کلاس پیش‌بینی شده		
		مشابه	نسبتاً مشابه	متفاوت
کلاس واقعی	مشابه	۱۸۷	۴۳	۰
	نسبتاً مشابه	۸	۳۲۶	۱۶
	متفاوت	۰	۱۴	۱۶۶

در جدول ۴-۱۱ نتایج کارایی روش پیشنهادی در پیکره فارسی PAN نمایش داده شده است. همانگونه که در جدول درهم‌ریختگی نیز به وضوح مشخص است به علت عدم توازن در داده‌های سه کلاس، معیار کارایی داده‌ها عدد نسبتاً مناسبی دریافت نکرده است. چالش‌ها دیگری که در آزمایشات با آنها برخورد داشتیم، اعمال تغییرات معنایی و ساختاری در اسناد مشکوک و همچنین کوچک بودن بخش به سرقت رفته نسبت به کل سند مشکوک بود. چالش اول باعث کاهش کارایی الگوریتم در تشخیص اسناد به علت عدم دخالت روابط معنایی بین واژگان می‌شود. چالش دوم نیز نقاط مرزی بین کلاس نسبتاً مشابه و متفاوت ایجاد می‌نماید که خود باعث کاهش درجه صحت و بازخورد در نتایج می‌باشد.

جدول ۴-۱۱ نتایج کارایی روش پیشنهادی در پیکره PAN

کلاس	*Recall*	*Precision*	*F-Measure*
مشابه	۰/۷۴	۰/۶۹	۰/۷۱
نسبتاً مشابه	۰/۵۵	۰/۹۵	۰/۶۹
متفاوت	۱	۰/۵۰	۰/۶۶

جدول ۴-۱۲ شامل نتایج کارایی روش پیشنهادی در پیکره ارزیابی شباهت آزمایشگاه وب‌کاوی دانشگاه صنعتی شاهرود می‌باشد. در این بخش به علت تفکیک شده بودن اطلاعات و همچنین عدم دخالت روابط معنایی در جفت اسناد پیکره، واضح است که سیستم به خوبی مشابهت‌های ظاهری را تشخیص داده است و توانسته نمره بسیار مناسبی در تمامی معیارها دریافت کند.

جدول ۴-۱۲ نتایج کارایی روش پیشنهادی در پیکره ارزیابی شباهت بین عبارات

کلاس	*Recall*	*Precision*	*F-Measure*
مشابه	۰/۸۱	۰/۹۵	۰/۸۷
نسبتاً مشابه	۰/۹۳	۰/۸۶	۰/۸۹
متفاوت	۰/۹۲	۰/۹۱	۰/۹۱

۴-۶- مقایسه روش پیشنهادی با کارهای مرتبط روی پیکره PAN

متاسفانه کارهای زیادی در زبان فارسی و بر روی پیکره PAN فارسی در دسترس نیست. ممتاز و همکاران [۴۲] روش مبتنی بر گراف را بر روی پیکره PAN فارسی پیاده‌سازی کرده‌اند که با استفاده از ارتباطات معنایی بین کلمات و رخداد هم‌زمان آنها شباهت بین دو سند را محاسبه می‌کند. در این روش از ۳۰۰ سند مشکوک و ۱۵۰۰ سند مرجع از پیکره PAN، برای کشف سرقت علمی در زبان فارسی استفاده شده‌است. راهکار ارزیابی در این پژوهش به این صورت است که اسناد به دو کلاس مشابه و متفاوت تقسیم می‌شوند. برای مقایسه مدل پیاده‌سازی شده در این پژوهش با کارهای مرتبط، دو کلاس مشابه و نسبتاً مشابه را به عنوان یک کلاس مثبت و کلاس متفاوت را به عنوان یک کلاس منفی درنظر می‌گیریم. بنابراین با در نظر گرفتن دو کلاس و محاسبه *Recall* و *Precision* بر مبنای دو کلاس به وجود آمده، صحت و بازخوانی

جدید حاصل خواهد شد. در جدول ۴-۱۳ مقایسه روش پیشنهادی و روش مبتنی بر گراف نمایش داده شده است. نتایج حاصل از این جدول حاکی از آن است که معیار پیشنهادی، علی‌رغم در نظر نگرفتن معنا در محاسبه شباهت، اما باز هم از کارایی بسیار خوبی نسبت به سایر روش‌ها برخوردار می‌باشد.

جدول ۴-۱۳ مقایسه کارایی روش پیشنهادی و روش مبتنی بر گراف

روش	*Recall*	*Precision*	*F-Measure*
روش پیشنهادی	۰/۸۳۲	۰/۷۴۹	۰/۷۸۷
روش مبتنی بر گراف	۰/۸۳۹	۰/۷۵۶	۰/۷۹۵
روش n-گرام	۰/۵۵۶	۰/۵۷۵	۰/۵۶۵

۴-۷- نتیجه‌گیری

در این فصل پس از معرفی پیکره مورد استفاده برای محاسبه شباهت و معیارهای ارزیابی، نتایج آزمایشات ارائه گردید. در مقایسه عبارات کوچک معیار جارو-وینکلر برای محاسبه شباهت بین گرام‌ها بهتر عمل می‌کند. همچنین نتایج مربوط به آزمایشات روی هر یک از پیکره‌ها در جداول جداگانه‌ای نمایش داده شد. روش پیشنهادی با دو روش دیگر مقایسه گردید که این نتایج نشان از در نظر گرفتن ویژگی‌های مطلوب برای محاسبه شباهت دارد.

فصل پنجم

بحث و نتیجه گیری

۵- بحث و نتیجه‌گیری

۵-۱- خلاصه تحقیق

در این پژوهش یک روش محاسبه شباهت بین متون مبتنی بر تکنیک‌های لغوی ارائه گردید. برای استخراج ویژگی‌ها از دو روش مبتنی بر هش و روش دگردیسی‌های کلی جملات استفاده نموده‌ایم. از طرفی نحوه شناسایی محتوای مشابه را در محیط غیر قطعی و با استفاده از سیستم استنتاج فازی پیاده‌سازی نمودیم. در نتیجه روشی ارائه گردید که سعی بر کم کردن ابهام و بالا بردن درجه اطمینان نسبت به سنجش دارد. این روش شامل سه بخش پیش‌پردازش، قطعه بندی و استخراج ویژگی‌ها و سیستم استنتاج شباهت‌سنجی لغوی است. در نهایت با معرفی ۵ ویژگی قواعد فازی توسط متخصصین استخراج گردید. خروجی سیستم استنتاج به صورت متغیرهای زبانی در سه کلاس متفاوت، نسبتاً مشابه و مشابه دسته‌بندی می‌شود که خود از نوآوری‌های این سیستم محسوب می‌شد.

برای ارزیابی سیستم فوق ما از پیکره PAN فارسی و پیکره ارزیابی شباهت بین عبارات استفاده نمودیم که نتایج آن در فصل چهارم به تفصیل بیان شد. نتایج نشان‌دهنده آن است که کارایی و دقت روش پیشنهادی، علی‌رغم در نظر نگرفتن بخش‌های معنایی متون ورودی، از کیفیت مطلوب برخوردار است. این روش در بیش از ۷۷ درصد موارد، تحلیل درستی از شباهت دو سند ارائه می‌هد.

۵–۲- پیشنهاد برای کارهای آینده

گرچه این کار تحقیقاتی پایان یافته است، اما هدف کلی ایجاد یک پلتفرم سریع و قابل دسترسی برای محاسبه شباهت و سرقت علمی-ادبی در زبان فارسی بوده است. در تمام مراحل سعی شده است تا سیستم به صورت کارا قابل استفاده مجدد باشد تا زمینه تحقیق پژوهشگران در این حوزه باشد. از این رو نقایص کدها و الگوریتم‌های این تحقیق در آینده برطرف خواهند شد. از جمله پیشنهادات برای ادامه‌ی کار عبارت‌اند از:

۱. طراحی یک ریشه‌یاب موثر در زبان فارسی و کاربردی برای ریشه‌یابی افعال و کلمات.
۲. توسعه ایست‌واژه‌ها و کلمات پرکاربرد قابل حذف از متون در زبان فارسی با استانداردهای پردازش زبان مورد استفاده همگانی.
۳. بهبود روش تشخیص بخش مشابه مبنی بر جمله و ارائه یک الگوریتم خوشه‌بندی برای تشخیص بخش‌های مشابه بین دو متن و خارج شدن از شکاندن متون به جملات به عنوان واحد ارزیابی مشابهت.
۴. توسعه سیستم استنتاج فازی و افزودن ویژگی‌هایی همانند در نظر گرفتن معنای کلمات و استفاده از پیکره‌های آنلاین همچون خزنده‌ها و ویکی‌پدیا برای محاسبه شباهت بین دو سند.

٦- مراجع

1. Gomaa, W.H. and A.A. Fahmy, *A survey of text similarity approaches.* International Journal of Computer Applications, 2013. pp. 13-18.
2. Alzahrani, S. and N. Salim, *Fuzzy semantic-based string similarity for extrinsic plagiarism detection.* Braschler and Harman, 2010: pp. 1-8.
3. Khoshnavataher, K., et al., *Developing monolingual Persian corpus for extrinsic plagiarism detection using artificial obfuscation.* Notebook for PAN at CLEF, 2015.
4. Singh, S., *Statistical Measure to Compute the Similarity between Answers in Online Question-Answering Portal.* International Journal of Computer Applications, 2014. **103**.
5. Mihalcea, R., C. Corley, and C. Strapparava. *Corpus-based and knowledge-based measures of text semantic similarity*. in *AAAI*. 2006.
6. Hall, P.A .and G.R. Dowling, *Approximate string matching.* ACM computing surveys (CSUR), 1980. **12**(4): pp. 381-402.
7. Jaro, M.A., *Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida.* Journal of the American Statistical Association, 1989. **84**(406): pp. 414-420.
8. Winkler, W.E., *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.* 1990.
9. Potthast, M., et al. *An evaluation framework for plagiarism detection*. in *Proceedings of the 23rd international conference on computational linguistics: Posters*. 2010. Association for Computational Linguistics.
10. منفرد, ح.ک.ا.ح., بهبود تعیین شباهت معنایی دو کلمه با استفاده از میانگین شباهت معنایی. پنجمیم کنفراوس داده کایی ایران، دانشگاه صنعتی امیرکبیر، ۱۳۹۰.
11. Danielsson, P.-E., *Euclidean distance mapping.* Computer Graphics and image processing, 1980. **14**(3): pp. 227-248.
12. اعتمادی, ر.ا., ن. مقدم چرکری، م. صالح نمدی, الگوریتمی جهت خوشه بندی صفحات وب بر اساس محتوای داده ای، سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران. ۱۳۸۶، انجمن کامپیوتر.
13. Miller, G.A., et al., *Introduction to WordNet: An on-line lexical database.* International journal of lexicography, 1990. **3**(4): pp. 235-244.

14. Gruner, S. and S. Naven. *Tool support for plagiarism detection in text documents*. in *Proceedings of the 2005 ACM symposium on Applied computing*. 2005. ACM.
15. Zechner, M., et al. *External and intrinsic plagiarism detection using vector space models*. in *Proc. SEPLN*. 2009.
۱۶. راکیان, ش.، ف. صافی ا صفهانی، ح. ر ستگاری، برر سی رو شهای ت شخیص سرقت علمی - ادبی بیرونی درمحیط تک زبانه، اولین همایش ملی رویکردهای نوین در مهندســـی کامپیوتر و بازیابی اطلاعات. ۱۳۹۲، دانشگاه آزاد اسلامی واحد رودسر و املش.
17. Wallach, H.M. *Topic modeling: beyond bag-of-words*. in *Proceedings of the 23rd international conference on Machine learning*. 2006. ACM.
18. Heintze, N. *Scalable document fingerprinting*. in *1996 USENIX workshop on electronic commerce*. 1996.
19. Stein, B. *Principles of hash-based text retrieval*. in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007. ACM.
20. Wibowo, A.T., K.W. Sudarmadi, and A.M. Barmawi. *Comparison between fingerprint and winnowing algorithm to detect plagiarism fraud on Bahasa Indonesia documents*. in *Information and Communication Technology (ICoICT), 2013 International Conference of*. 2013. IEEE.
21. Stein, B., S.M. zu Eissen, and M. Potthast. *Strategies for retrieving plagiarized documents*. in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007. ACM.
22. Meuschke, N. and B. Gipp, *State-of-the-art in detecting academic plagiarism.* International Journal for Educational Integrity, 2013. p. 9.
۲۳. اســکندری، س.ز. ب. مینایی بیدگلی، بهبود کارایی الگوریتم شــیگلینگ با اســتفاده از شــینگلهای وزن دار، پانزدهمین کنفرانس بین المللی سالانه انجمن کامپیوتر ایران. ۱۳۸۸, انجمن کامپیوتر.
24. Hoad, T.C. and J. Zobel, *Methods for identifying versioned and plagiarized documents.* Journal of the Association for Information Science and Technology, 2003. **54**(3): pp. 203-215.
25. Schleimer, S., D.S. Wilkerson, and A. Aiken. *Winnowing: local algorithms for document fingerprinting*. in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*. 2003. ACM.
26. Pataki, M. *Plagiarism detection and document chunking methods*. in *Proceedings of the twelfth international conference on World Wide Web, WWW2003. Budapest*. 2003. NIIF.
27. Ehsan, N. and A. Shakery. *A Pairwise Document Analysis Approach for Monolingual Plagiarism Detection*. in *FIRE (Working Notes)*. 2016.

28. Manku, G.S., A. Jain, and A. Das Sarma. *Detecting near-duplicates for web crawling*. in *Proceedings of the 16th international conference on World Wide Web*. 2007. ACM.

۲۹. کامران، ک، ع. احمدی، ف. کاظمی و نهری، کشـــف ســـرقت ادبی در متون فارســـی به کمک الگوریتمهای اثرانگشت، بیست و یکمین کنفرانس مهندسی برق ایران. ۱۳۹۲، دانشگاه فردوسی مشهد.

30. Sadowski, C. and G. Levin, *Simhash: Hash-based similarity detection*. 2007.

31. Arasu, A., V. Ganti, and R. Kaushik. *Efficient exact set-similarity joins*. in *Proceedings of the 32nd international conference on Very large data bases*. 2006. VLDB Endowment.

32. Jiang, Y., et al., *String similarity joins: An experimental evaluation.* Proceedings of the VLDB Endowment, 2014. **7**(8): pp. 625-636.

۳۳. آهنگربهان، ح. غ. منتظر, مدل سازی عدم قطعیت در سنجش شباهت لغوی محتوای منابع وب فارسی، اولین کنفرانس بین المللی وب پژوهی. ۱۳۹۴، دانشگاه علم و فرهنگ.

34. Baždarić, K., *Plagiarism detection–quality management tool for all scientific journals.* Croatian medical journal, 2012. **53**(1): pp. 1-3.

35. Rutkowski, L. and K. Cpalka, *Flexible neuro-fuzzy systems.* IEEE Transactions on Neural Networks, 2003. **14**(3): pp. 554-574.

۳۶. غلامعلی، س.ل.و.م، طراحی و پیاده سازی سیستم هوشمند شناسایی رفتار مشکوک در بانکداری الکترونیک به کمک نظریه مجموعه‌های فازی. فصلنامه فناوری اطلاعات و ارتباطات ایران، ۱۳۸۷.

37. Metzler, D., et al. *Similarity measures for tracking information flow*. in *Proceedings of the۱۴ th ACM international conference on Information and knowledge management*. 2005. ACM.

38. Zou, D., W.-J. Long, and Z. Ling. *A Two-Phase Plagiarism Detection Method*. in *Internet Technology and Applications (iTAP), 2011 International Conference on*. 2011. IEEE.

39. Alzahrani, S.M., *Statement-Based Fuzzy-Set Information Retrieval versus Fingerprints Matching for Plagiarism Detection in Arabic Documents*.

40. Ekbal, A., S. Saha, and G. Choudhary. *Plagiarism detection in text using Vector Space Model*. in *Hybrid Intelligent Systems (HIS), 2012 12th International Conference on*. 2012. IEEE.

41. Mahmoodi, M. and M.M. Varnamkhasti, *Design a Persian Automated Plagiarism Detector (AMZPPD).* arXiv preprint arXiv:1403.1618, 2014.

۴۲. ممتاز, م.، م. صالحی، ه. ویسی, تشخیص سرقت علمی اسناد فارسی با رویکرد مبتنی بر گراف، هشـــتمین کنفرانس بین المللی فناوری اطلاعات و دانش. ۱۳۹۵, انجمن فناوری اطلاعات و ارتباطات ایران.

۴۳. آهنگربهان, ح، غ. منتظر, شباهت سنجی لغوی در وب گاه های تخصصی فارسی به کمک سیستم نروفازی، دومین کنفرانس بین المللی وب پژوهی. ۱۳۹۵, دانشگاه علم و فرهنگ.
44. Paul, M. and S. Jamal, *An improved SRL based plagiarism detection technique using sentence ranking.* Procedia Computer Science, 2015. **46**: pp. 223-230.
45. Ceska, Z. and C. Fox. *The influence of text pre-processing on plagiarism detection*. 2011. Association for Computational Linguistics.
46. Davarpanah, M.R., M. Sanji, and M. Aramideh, *Farsi lexical analysis and stop word list.* Library Hi Tech, 2009. **27**(3): pp. 435-449.

Abstract

Creating a system that can effectively determine the distance between two expressions has been a concern of researchers in artificial intelligence and data retrieval. The similarity measure of the two terms are used in a wide range of applications, such as natural language processing, query correction, semantic spelling, plagiarism detection, document comparisons and other areas of data retrieval. The input of detecting systems can be a text or a set of documents and texts, and the output is the result of a system's judgment about the similarity of the input sentences. Ultimately, the proximity of the judgment of the system to human judgment is indicative of the good functioning of the system.

In this paper, we propose a method for calculating the uncertainty-based lexical similarity. To achieve this purpose, we use fingerprinting-based algorithms and Winnowing techniques as similarity calculation measures. Also, by considering several parameters to calculate the number of phrases, we use the fuzzy inference system to decide on uncertainty conditions. In this method, the characteristics are based on the technique of hashing and general transformation of sentences that these criteria have a good speed and accuracy. After calculating the similarity of the two inputs, similar sentences are extracted by the system. The method has been evaluated on the PAN database and the Shahrood University of Technology Similarity assessment, which with a precision of 78%, separates the documents into three different, relatively similar and similar classes.

Keywords: lexical similarity, plagiarism, fuzzy inference system, fingerprint, winnowing

Shahrood University of Technology

Faculty of computer engineering

Thesis submitted in partial fulfillment of the requirement for the degree of Master of Science in Computer Engineering in the subject of Artificial Intelligence

Defining a Fuzzy Distance Function to Compare Expressions

By:

Mehrdad Kaheh

Supervisor:

Dr. M. Zahedi

August 2017