مدل‌سازی عدم‌قطعیت در سنجش شباهت لغوی محتوای منابع وب فارسی

حمید آهنگربهان[۱]، غلامعلی منتظر[۲]

[۱] دانشگاه تربیت مدرّس، دانشکدهٔ فنی و مهندسی، گروه مهندسی فنّاوری اطلاعات،
h.ahangarbahan@modares.ac.ir

[۲] دانشگاه تربیت مدرّس، دانشکدهٔ فنی و مهندسی، گروه مهندسی فنّاوری اطلاعات،
montazer@modares.ac.ir

چکیده

در دنیای کنونی، کاربران به راحتی می‌توانند با رونوشت‌برداری از منابع وبی، سند و نوشته‌ای بدون ارجاع به مرجع اصلی به نام خود ارائه دهند که این عمل مصداقی از دستبرد ادبی است. تاکنون روش‌ها و سنجه‌های مختلفی در سامانه‌های دستبرد ادبی برای ارزیابی مشابهت دو سند و تشخیص دستبرد ادبی ارائه شده که تنها به صورت قطعی میزان شباهت بین دو متن را سنجیده و به نوع محتوای متون چندان توجهی نمی‌کردند. در این مقاله با توجه به کیفی بودن عوامل اثرگذار بر سنجش شباهت بین دو متن، روش جدیدی با استفاده از نظریهٔ مجموعه فازی ارائه شده است. در این روش ابتدا، واژه‌های هر متن به دو دسته عمومی و حوزه‌محور(تخصصی) بخش‌بندی شده و سپس برای هر دسته سنجه‌ها و ویژگی‌ها متناسب آن مستخرج‌شده و در نهایت با استفاده از سیستم استنتاج فازی، میزان شباهت لغوی بین دو محتوای صفحه وب ارزیابی می‌شود. روش مذکور بر روی مقاله‌های یادگیری الکترونیکی مورد ارزیابی قرار گرفته که با دقت بیش از ۷۵٪ امکان شناسایی محتوای اسناد مشابه را داراست و به همین دلیل توانمندی لازم برای استفاده در حوزه شناسایی محتوای منابع وبی مشابه و همین‌طور تشخیص دستبرد محتوایی را داراست.

کلمات کلیدی

دستبرد ادبی، متن فارسی، نظریهٔ مجموعه فازی،سنجه شباهت‌سنجی، شباهت‌سنجی لغوی

۱- مقدمه

با گسترش روزافزون اطلاعات بر روی وب[۱]، دستیابی به انواع اطلاعات به‌راحتی امکان‌پذیر شده است و از طرفی این حجم وسیع اطلاعات اگرچه مزایایی همچون دسترسی سریع، آسان و متنوع را فراهم آورده ولی چون ارسال اطلاعات در وب از مرجع خاصی صورت نمی‌پذیرد، اطلاعات تکراری و مشابه فراوانی در این محیط وجود داشته و به نوعی باعث سردرگمی و اتلاف وقت محققان و کاربران خواهد شد. از طرفی استفاده نادرست از این اطلاعات و مطالب و ارائه آن به نام خود، معضل دیگری را به وجود آورده که روز به روز در حال افزایش است؛ این عمل که مصداقی از دستبرد ادبی[۲] بوده به شدت محیط‌های دانشگاهی را تهدید می‌کند. ارائه راهکارهای که بتواند دستبرد ادبی، اسناد مشابه و تکراری را شناسایی کند از حوزه‌های مورد علاقه محققان در سال‌های اخیر بوده و در مسائل فراوانی همچون ترجمه ماشینی[۳]، ابهام‌زدایی حس واژه[۴]، خلاصه‌سازی[۵]، تشخیص دگرمعنانویسی[۶] و خوشه‌بندی متون[۷] کاربرد دارد. به عنوان مثال، شباهت‌سنجی[۸] دقیق متن باعث عملکرد بهتر در تشخیص دگرمعنانویسی شده و متون را با کارایی بهتری می‌توان خوشه‌بندی و از نگهداشت اطلاعات اضافه جلوگیری کرده و هزینه‌های انباره‌های اطلاعاتی را به شدت کاهش داد[2] و یا به راحتی تحقیق درست، با ارجاع منابع اطلاعاتی و وبی را می‌توان شناسایی کرد.

تاکنون پژوهش‌های فراوانی برای تشخیص اسناد مشابه و همچنین دستبرد ادبی انجام گرفته است. به عنوان نمونه، الظهرانی و همکارانش و همچنین عثمان و همکارانش مرور کاملی از روش‌های که تاکنون برای شناسایی اسناد مشابه و مقابله با دستبرد ادبی پیشنهاد شده و همچنین دسته‌بندی انواع آن ارائه کرده‌اند[3,4].

در تحقیقات و روش‌های که تاکنون برای شناسایی اسناد مشابه و تشخیص دستبرد ادبی انجام و ارائه شده، نحوهٔ سنجش شباهت بین دو واژه (عبارت یا جمله) که تاثیر عمده‌ای در دقت تشخیص داشته و متاثر از بیان خبرگان چه به صورت بیان مستقیم و یا غیر مستقیم(از طریق پایگاه دانش) بوده چندان به صورت کیفی مورد تحقیق قرار نگرفته و در اکثر این روش‌ها به صورت کمّی(قطعی) بیان شده است. در نتیجه با توجه اینکه اکثر منابع وبی از نوع علمی و تخصصی بوده، به کارگیری رویکردهای قبلی در این محتوا سنجش شباهت را دچار محدودیت ابهام کرده و کارایی آنها را کاهش می‌دهد [5,6]. نظریهٔ مجموعه فازی از جمله نظریهٔ‌های ریاضی است که در ارائه راه‌حل برای مسائل مبهم کارا بوده است به همین منظور در این مقاله از این نظریهٔ استفاده شده وروشی برای برطرف‌سازی نقص اطلاعاتی دربارهٔ سنجش شباهت در محتوای تخصصی و علمی منابع وبی با استفاده بیان فازی ارائه گشته که با دقت و اطمینان بالایی شباهت لغوی دو محتوا را محاسبه و در نهایت دستبرد ادبی و اسناد مشابه را تشخیص می‌دهد.

ادامهٔ متن این مقاله به صورت این صورت تنظیم شده است که در بخش ۲ مسئله تشخیص دستبرد ادبی و شناسایی اسناد مشابه تعریف شده و در بخش ۳ کلیات نظریهٔ مجموعه فازی بیان و در بخش ۴ معماری روش شباعت‌سنجی لغوی به صورت جزیی آورده شده است. در بخش ۵ نتایج پیاده‌سازی روش طراحی شده بیان می‌شود و در نهایت بخش ۶ به نتیجه‌گیری می‌پردازد.

۲- مسئله دستبرد ادبی(محتوایی)

دستبرد ادبی(محتوایی) را می‌توان در ساده‌ترین حالت به صورت «رونوشت‌برداری از اسناد یا برنامه بدون تأیید مرجع اصلی دانست که یکی از شایع‌ترین موارد آن، استفاده از مطالب دیگران در نوشته‌ای به نام خود است» [7]. به لحاظ ریاضی این تعریف را می‌توان چنین بیان کرد [8]:

مجموعه داده‌ای اسناد مشکوک D_q و مجموعهٔ مرجع D را در نظر بگیرید. تشخیص دستبرد ادبی عبارت است از : یافتن بخش‌های مشکوک s_q از D_q که شبیه به بخش‌های s_x از یکی از اسناد موجود در D.

مهمترین نکته در حل هر مسئله شناخت درست آن مسئله و به دست آوردن فضای حالت آن است. از طرفی از آنجا که معمولاً کاربران از تغییر ترتیب واژه‌ها و یا جابه‌جا کردن بخش‌های مختلف متن برای بازنویسی متن (به خصوص در متون تخصصی) استفاده می‌کنند[9]، تشخیص این نوع محتوای مشابه و دستبردی دشوار خواهد بود و نیاز است برای به دست آوردن کارایی بهتر بیان دقیق‌تری از مسئله ارائه و عوامل تاثیرگذار بر سنجش و تشخیص شباهت بین دو متن شناسایی شود. در نتیجه در این مسئله باید به دنبال فضای حالت پرسش زیر باشیم:

«دو متن تخصصی به زبان فارسی در اختیار داریم چگونه می‌توان شباهت بین این دو متن را با دقت و اطمینان فراوان سنجید»

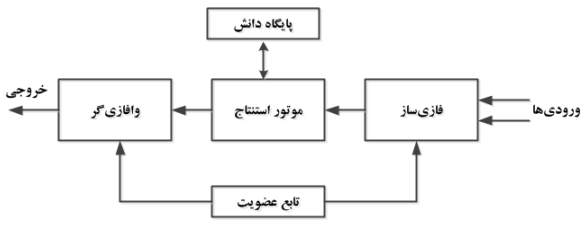
در حال حاضر از رویکردهای متفاوتی همانند مبتنی بر بردار[۹] و یا پایگاه دانش‌محور[۱۰] با منابع مختلف برای محاسبه میزان شباهت بین دو واژه استفاده می‌شود[10,11]. وجود این رویکردها نشان‌دهنده آن است شباهت مفهومی ذهنی، مبهم و نامعلوم است و نیاز است بیان شفاف‌تری از شباهت ارائه گردد.

تلاش‌های اندکی در حوزه شباهت‌سنجی بر روی بیان شفاف نحوه سنجش، عدم‌دقت و عدم‌اطمینان منابع مورد استفاده برای آن انجام شده است. از جمله این تلاش‌ها می‌توان به کارهای الظهرانی و سلیمی[12] که سنجه شباهت‌سنجی بین دو واژه را به صورت تابع عضویت فازی مدل کرده و گوپتا و همکاران[13] که با بازتعریف تابع عضویت فازی کار الظهرانی و سلیمی در بازه‌های جزیی‌تر، شباهت بین اسناد PAN 2012 را سنجیده و نشان دادند تعریف دقیق‌تر تابع عضویت فازی سنجهٔ شباهت‌سنجی می‌تواند برای شباهت‌سنجی مفید است؛ اشاره کرد.

در این تحقیقات هیچ‌گونه تمایزی برای متن مورد بررسی در نظرگرفته نشده و همهٔ واژه‌های متن به دید یکسانی ملاحظه شده و میزان اثرگذاری آنها بر شباهت و همچنین نظرهای خبرگان نسبت به نحوه سنجش دیده نشده است. از طرفی تنها یک سنجه شباهت‌سنجی برای سنجش در نظر گرفته شده که چندان سنجش کاملی را به دست نمی‌دهد. در ادامه کلیات نظریهٔ مجموعه فازی و روش پیشنهادی برای محاسبه شباهت بین دو متن را تشریح خواهد شد.

۳- نظریهٔ مجموعه فازی

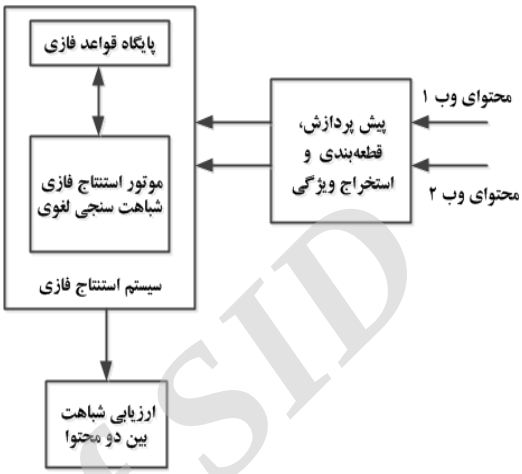
عدم‌قطعیت یکی از مواردیست که برای انسان‌ها در شناخت محیط به وجود می‌آید. انسان‌ها معمولاً در تحلیل و بیان کمّی دچار مشکل بوده و از سوی دیگر در قضاوت کیفی در مورد به صورت کارا موفق عمل می‌کنند. زاده برای حل این مشکل نظریهٔ مجموعه فازی را ارائه داد. این نظریهٔ، چارچوب مناسبی را برای محاسبهٔ داده‌ها و اطلاعات غیرقطعی[۱۱] و مبهم[۱۲] ارائه می‌کند. همچنین این نظریه می‌تواند روابط موضوعی و عدم‌قطعیت را به صورت ریاضی بیان کند. سیستم استنتاج فازی نیز بخش تصمیم‌گیرنده و مغز سیستم فازی است که فرایند نگاشت ورودی به خروجی را با استفاده از نظریهٔ مجموعه‌های فازی انجام می‌دهد.[14-16]. شکل (۱) نمونه‌ای از که سیستم استنتاج فازی که شامل ۵ بخش فازی‌ساز،تابع عضویت، موتور استنتاج، پایگاه دانش و وافازی‌گر است را نشان می‌دهد. قلب هر سیستم فازی «پایگاه دانش» است از ترکیب دانش خبرگان حوزهٔ مورد بحث و به شکل قواعدی از متغیرهای زبانی تشکیل می‌شود. در این چارچوب استنتاج از طریق مجموعه‌ای از قواعد «اگر–آنگاه» انجام می‌شود که هر یک از این قواعد به کمک مجموعه‌های فازی تعریف می‌شوند.[۱].



شکل(۱): چهارچوب سیستم استنتاج فازی

٤- معماری روش شباهت‌سنجی لغوی

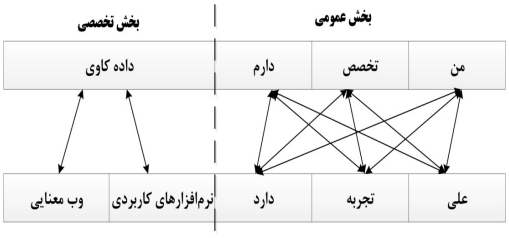
شکل (۲) معماری روش شباهت‌سنجی لغوی برای محتوای دو منبع وبی را نشان می‌دهد. این معماری شامل سه جزء اصلی پیش‌پردازش، قطعه‌بندی و استخراج ویژگی و در نهایت سیستم استنتاج فازی است. نحوه کارکرد هر یک از این اجزا در ادامه خواهد شد.



شکل(۲): معماری روش شباهت‌سنجی لغوی

٤-۱- پیش‌پردازش و قطعه‌بندی

در این بخش، با توجه به این‌که محتوای هر سند وبی شامل واژه‌های اصلی و غیر اصلی همانند ایست‌واژه است در مرحلهٔ ابتدایی واژه‌های غیر اصلی از محتوای سند حذف می‌گردد. از طرفی از آنجا که فعل‌ها به صورت مختلف در سند ظاهر می‌گردد در مرحلهٔ بعد این واژه‌ها ریشه‌یابی می‌شوند. در نهایت متن پالایش شده به بخش محتوای تخصصی‌محور و عمومی تقسیم می‌شود. باید توجه داشت در محتوای تخصصی هر واژه با وزن‌های مختلف به معنای متفاوت ارتباط دارند. در نتیجه تشخیص معنای هر واژه برای سنجش بهتر و بااطمینان‌تر نیاز است و باید از منابع الکترونیکی خاص همان محتوا استفاده کرد. به همین منظور در روش پیشنهادی هر متن به دو بخش تخصصی و عمومی تقسیم‌بندی شده است. در بخش عمومی واژه‌های عمومی همانند واژه «کتاب» که در اکثر محتواها استفاده شده و در بخش تخصصی واژه‌های تخصصی یک حوزه خاص همانند واژه «یادگیری الکترونیکی » قرار می‌گیرد. برای نمونه دو جمله «من در داده‌کاوی تخصص دارم» و «علی در نرم‌افزارهای کاربردی و وب‌معنایی تجربه دارد» در نظر بگیرید. این جمله‌ها با استفاده از هستان‌نگار[۱۳] تخصصی مطابق با شکل (۳) به دو بخش عمومی و تخصصی تقسیم می‌شوند.



شکل(۳): نحوه مقایسه زوج‌واژه‌ها در زوج جمله‌ها

٤-۲- استخراج ویژگی و سنجه شباهت‌سنجی

پس از بخش‌بندی هر متن، آنگاه بخش‌های مرتبط با هم مقایسه می‌گردد به این صورت که بخش تخصصی متن اول با بخش تخصصی متن دوم و همین‌طور برای بخش عمومی هر دو متن نیز مقایسه انجام می‌گیرد. در روش پیشنهادی از سنجه‌ها و ویژگی‌های رویکرد شباهت‌سنجی لغوی[۱۴]، استفاده شده است. این ویژگی‌ها و سنجه‌ها در ادامه معرفی می‌شوند.

۴-۲-۱- ویژگی اختلاف تعداد واژه‌ها در هر بخش

این ویژگی به آن علت در نظر گرفته شده است که معمولا متونی که از منظر لغوی با هم مشابه باشند تا حدودی زیادی اختلاف واژه کمی با هم دارند. اگرچه این ویژگی، یک ویژگی عام نیست ولی به نوعی مکمل برای شباهت‌سنجی است.

۴-۲-۲- معیار F

این معیار تعداد واژه‌های مشترک بین دو متن را از طریق تولید n-گرم‌ها[۱۵] در سطح (واژه یا کاراکتر) به ترتیب تقسیم بر طول متن اول و دوم کرده و سپس از این دو نسبت میانگین هندسی می‌گیرد. این معیار، مقایسه متن اول با متن دوم به صورت زیر انجام می‌دهد[17].

۴-۲-۳- سنجه پرش-گرم[۱۶]

این سنجه همانند n-گرم است با این تفاوت که می‌تواند برای تولید گرم‌ها در سطح واژه پرشی به اندازه n داشته باشد. طبق تحقیقات انجام گرفته تعداد کارای n عدد سه و چهار بوده است. این سنجه به نوعی ویژگی ترتیب و همچنین وجود عبارت‌های چندواژه‌ای مشترک در هر بخش را در نظر می‌گیرد.

جدول (۱) فهرست کاملی از سنجه‌ها و ویژگی‌های انتخاب شده با توجه به رویکرد استفاده شده ارائه می‌دهد. با توجه به این سنجه‌ها و ویژگی‌ها می‌توان متغیرهای بخش استنتاج فازی را طراحی نمود.

جدول (۱): ویژگی‌ها و سنجه‌های انتخاب شده

نوع واژه (بخش‌بندی)	
عمومی	تخصصی
۱- اختلاف تعداد واژه‌های عمومی	۱-اختلاف تعداد واژه‌های تخصصی
۲- سنجه شباهت‌سنجی پرش-گرم	۲-سنجه شباهت‌سنجی پرش-گرم
۳- معیار F	۳-معیار F

٤-۳- استنتاج فازی

پس از تعریف سنجه‌ها و ویژگی‌های مناسب، از آنجا که سنجش شباهت بین دو متن تخصصی عملی است که با توجه به دانش افراد و خبرگان 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 نامعلوم‌بودن سنجش بود. این نحوه تعریف فازی از ارزیابی شباهت در جدول (۲) نمایش داده شده است.

جدول (۲) : متغیر زبانی شباهت

متغیر زبانی	بازه عددی
متفاوت (Different)	(−∞ ,۰ ,۰/۱ ,۰/۳۵)
تا حدودی مشابه(Semi Similar)	(۰/۱,۰/۳۵ ,۰/۵۵ ,۰/۷۵)
مشابه(Similar)	(۰/۶۵ ,۰/۷۵,۱,+∞)

از آنجا که ارزیابی شباهت بین دو متن، فازی تعریف شده، سنجه‌های که در ورودی سیستم‌های استنتاج محاسبه می‌شوند نیز به صورت فازی مدل شده‌اند. چون بازه عددی و جنس برخی متغیرهای بکار گرفته‌شده؛ برای راحتی فهم در جدول (۳) انواع متغیرهای فازی و همچنین بازه عددی و متغیر زبانی آورده و در جدول (۴) متغیرهای استفاده شده در هر سیستم به همراه جنس آن مشخص شده است. باید توجه داشت برخی از این متغیرها به صورت تابع فازی برای هر بخش(تخصصی/عمومی) تعریف شده‌اند که در ادامه به صورت جزئی‌تر نحوه محاسبه آن تشریح می‌شود.

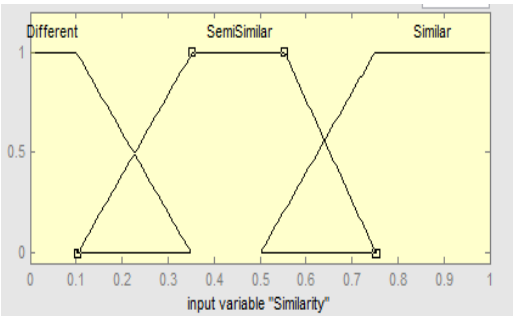
جدول (۳): متغیرهای زبانی در ارزیابی شباهت بین دو متن

نوع متغیر	حرف اختصاری	بازه عددی	متغیر زبانی
نسبت تعداد واژه‌های بخش تخصصی بر تعداد واژه‌های بخش عمومی	T1	(−∞ ,۰ ,۰/۱ ,۰/۲۵)	کم
		(۰/۱۵,۰/۳۵ ,۰/۵۵ ,۰/۸)	متوسط
		(۰/۵۰ ,۰/۷۵,۱,+∞)	زیاد
ارزیابی شباهت بین دو واژه/جمله	T2	(−∞ ,۰ ,۰/۱ ,۰/۳۵)	متفاوت
		(۰/۱,۰/۳۵ ,۰/۵۵ ,۰/۷۵)	تا حدودی مشابه
		(۰/۶۵ ,۰/۷۵,۱,+∞)	مشابه

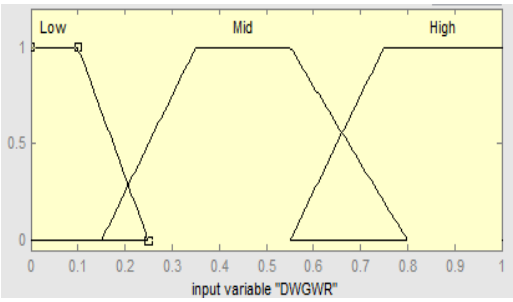
جدول (٤): متغیرهای سیستم فازی شباهت‌سنجی لغوی

بخش	نام متغیر	حرف اختصاری	جنس متغیر
عمومی	تابع شباهت پرش-گرم عمومی	V1	T2
	تابع شباهت معیار F عمومی	V2	T2
تخصصی	تابع شباهت پرش-گرم تخصصی	V3	T2
	تابع شباهت معیار F تخصصی	V4	T2
کل متن	نسبت تعداد واژه تخصصی به عمومی	V5	T1

شکل (۴) به نیز تابع عضویت برخی متغیرهای کاربردی در سیستم فازی را نشان می‌دهد.



ب- ارزیابی شباهت بین دو واژه/جمله



الف- نسبت تعداد واژه‌های‌بخش‌تخصصی‌بر تعداد واژه‌های بخش عمومی

شکل (٤): توابع عضویت متغیرهای ورودی و خروجی

۴-۳-۱- تابع فازی مقایسه زوج بخش

از آنجا که هر بخش شامل چندین واژه بوده برای مقایسه آنها با هم نیاز به مقایسه زوج واژه و سپس تجمیع نتایج آنها همانند آنچه برای مثال شکل (۲) خواهد بود. تنها تفاوت در این بخش به این صورت خواهد که مقایسه‌ها به صورت فازی خواهد بود و در نتیجه به عنوان نمونه ماتریس خروجی به صورت رابطه (۱) خواهد بود. درایه‌های این ماتریس هر کدام عدد فازی بوده که برای بدست آوردن نتیجه نهایی بین دو متن از ماکسیم به عنوان نرم S در هر سطر و ستون استفاده کرده و سپس مقادیر میانگین این دو عدد فازی را با استفاده از رابطه رزونفلد[۱۷] به دست خواهد آمد.

$$Sim_{ont}(S_1,S_2) = \left[\{(\text{مشابه},0.9),(\text{تاحدودی‌مشابه},0.2),(\text{متفاوت},0.05)\} \quad \{(\text{مشابه},0.7),(\text{تاحدودی‌مشابه},0.4),(\text{متفاوت},0.23)\}\right] \quad (۱)$$

برای بخش تخصصی همانند مثال مذکور محاسبات برای دو سنجه مبتنی بر هستان‌نگار عمل می‌شود و در نهایت برای هر سنجه خروجی فازی نهایی محاسبه می‌گردد که به عنوان ورودی سیستم‌های استنتاج فازی خواهد بود.

۴-۳-۲- پایگاه قواعد فازی

پس از تعریف دقیق متغیرهای فازی سیستم استنتاج قواعد فازی این سیستم با استفاده از مصاحبه با گروهی ۵ نفره از خبرگان زبان‌شناسی و مختصصان حوزه متن تخصصی استخراج شده است. برخی از این قواعد در جدول (۵) آورده شده است. برای واضح بودن متغیرهای شباهت‌سنجی ورودی سیستم‌ها به صورت کم، متوسط و زیاد بیان شده تا با خروجی سیستم تمایز داشته باشد.

جدول (۵): برخی قواعد فازی در سیستم فازی شباهت‌سنجی

شماره قاعده	نام متغیر					نتیجه سیستم
	V1	V2	V3	V4	V5	
۱	–	زیاد	–	کم	کم	مشابه
۲	–	زیاد	کم	کم	متوسط	تاحدودی مشابه
۳	–	زیاد	متوسط	کم	متوسط	تاحدودی مشابه
۴	–	زیاد	کم	کم	زیاد	متفاوت
۵	–	زیاد	–	متوسط	کم	مشابه
۶	کم	متوسط	کم	کم	متوسط	متفاوت
۷	متوسط	متوسط	کم	کم	متوسط	تاحدودی مشابه
۸	زیاد	متوسط	متوسط	کم	متوسط	مشابه
۹	–	متوسط	متوسط	کم	زیاد	تاحدودی مشابه
۱۰	–	کم	–	کم	–	متفاوت

در بخش نتایج عددی اجرای این سیستم فازی بر روی داده عددی نمایش داده می‌شود.

۵- ارزیابی روش

در این بخش نتایج عددی روش پیشنهادی را با استفاده از داده‌های واقعی گزارش خواهیم کرد. روش پیشنهادی در محیط‌های نرم‌افزاری Visual Studios .Net و با استفاده از پایگاه داده Microsoft SQL و نرم‌افزار متلب برای بخش سیستم فازی پیاده‌سازی شده و معماری روش به صورت شیئ‌گرا انجام گرفته که توانایی سنجش هر دو محتوای وبی را بر اساس سنجه‌های متفاوت و همچنین پایگاه دانش‌های مختلف داشته باشد.

از آنجا که هدف در این مقاله بررسی شباهت در متون تخصصی منابع وبی فارسی است و در حال حاضر هیچگونه پیکره یا مجموعه داده‌ای با این ویژگی در زبان فارسی و همین طور زبان‌های دیگر وجود ندارد. از مجموعه پایگاه مقاله‌های یادگیری الکترونیکی برای تولید پیکره استفاده شده است. این مجموعه داده شامل ۷۵۰ زوج‌جمله بوده که توسط تعدادی خبره به سه کلاس مشابه، تاحدودی مشابه و متفاوت دسته بندی شده است.

برای درک بهتر از روش پیشنهادی مراحل محاسبه شباهت را در این قسمت تشریح خواهیم کرد. جدول (۶) خروجی استنتاج سیستم فازی شباهت‌سنجی نشان می‌دهد. نتایج نهایی در این جدول به صورت پررنگ مشخص شده است.

برای ارزیابی عملکرد کارایی روش‌پیشنهادی از، شاخص‌های «دقت[۱۸]»، «بازخوانی[۱۹]» و «شاخص F» استفاده می‌شود. «دقت» و « بازخوانی» دو شاخص برای ارزیابی روش‌های بازیابی اطلاعات هستند. «بازخوانی»، میزان اسناد مرتبط بازیابی شده است و «دقت»، میزان اسناد بازیابی شدهٔ مشابه است و «شاخص F» میانگین هندسی دو شاخص دقت و بازخوانی است. اگر A نشان‌دهندهٔ مجموعهٔ جمله‌های بازیابی شده و R نشان‌دهندهٔ مجموعهٔ جمله‌های مشابه برای یک جملهٔ مفروض باشند، روابط دقت و بازخوانی به صورت رابطه (۲) و (۳) است[18]:

$$\Pr ecision = \frac{|A \cap R|}{|A|} \qquad (۲)$$

$$\mathrm{Re}\, call = \frac{|A \cap R|}{|R|} \qquad (۳)$$

جدول (٦): نتایج سیستم فازی

مورد #	متغیرهای فازی					نتیجه شباهت فازی		
	V1	V2	V3	V4	V5	مشابه	تاحدودی مشابه	متفاوت
۱	۰/۳۴	۰/۶۳	۰/۸۱	۰/۷	۰/۷۳	**۰/۸۲**	۰/۰۷	۰/۰۲
۲	۰/۴۲	۰/۵۷	۰/۹۳	۰/۶۱	۰/۱۶	**۰/۷٤**	۰/۲۳	۰/۱۴
۳	۰/۱۲	۰/۱۶	۰/۲۹	۰/۴۱	۰/۵۵	۰/۱۴	۰/۳۶	**۰/۷۷**

این دو شاخص در جهت عکس هم حرکت می‌کنند بدین معنا که بهبود یکی به کاهش دیگری منجر می‌شود. به منظور رفع این مشکل از میانگین متوازن آنها به صورت رابطه (۴) استفاده می‌شود[18]:

$$F - Measure = \frac{\mathrm{Re}\, call * \Pr ecision}{\mathrm{Re}\, call + \Pr ecision} \qquad (۴)$$

جدول (۷) نتایج ماتریس درهم‌ریختگی[۲۰] روش پیشنهادی را نشان می‌دهد. برای محاسبه معیارهای کارایی از آنجا که هدف در این تحقیق تشخیص دستبرد ادبی، دو کلاس «مشابه» و «تاحدودی‌مشابه» به عنوان کلاس مثبت[۲۱] و کلاس «متفاوت» به عنوان کلاس منفی[۲۲] در نظر گرفته شده است.

جدول (۷): ماتریس درهم‌ریختگی نتایج روش پیشنهادی

		کلاس پیش بینی شده		
		مشابه	تاحدودی‌مشابه	متفاوت
کلاس واقعی	مشابه	۱۱۳	۶۵	۷۲
	تاحدودی‌مشابه	۶۳	۱۴۴	۴۳
	متفاوت	۴۷	۶۲	۱۴۱

نتایج پیاده‌سازی واقعی روش پیشنهادی و همچنین مقایسه آن با پیاده‌سازی روش مقاله [13] و یک روش سنجه قطعی(تشابه کسینوسی) بر روی پیکره یادگیری الکترونیکی در جدول (۸) آورده شده است. نتایج نشان می‌دهد که روش پیشنهادی به میزان قابل قبولی بهتر از روش مقاله [13] و روش قطعی بوده و توانایی شناسایی محتوای وبی مشابه را دارد.

جدول (۸): نتایج کارایی روش پیشنهادی و با دو روش دیگر

روش شباهت‌سنجی لغوی	معیار F	دقت	بازخوانی
روش پیشنهادی	۰/۷۵	۰/۷۳	۰/۷۸
روش مقاله [13]	۰/۶۶	۰/۶۹	۰/۶۴
سنجه قطعی	۰/۶۴	۰/۶۷	۰/۶۳

٦- نتیجه‌گیری

در این مقاله، با توجه حجم گسترده محتوای وبی مشابه به دنبال ارائه راه‌حلی برای شناسایی این اسناد بوده‌ایم. از طرفی نحوه شناسایی محتوای مشابه در

روش‌های که تاکنون ارائه شده اکثراً به صورت قطعی بوده و عدم‌قطعیت در نحوه سنجش را در نظر نگرفته است. در نتیجه روشی ارائه شده که سعی بر کم‌کردن ابهام و همچنین بالابردن درجهٔ اطمینان نسبت به سنجش دارد و از نظریهٔ فازی که توانایی حل مسائل با عدم‌قطعیت را داراست، برای حل کمک گرفته شده است. روش طراحی‌شده شامل سه بخش پیش‌پردازش و قطعه‌بندی، استخراج ویژگی و سیستم استنتاج شباهت‌سنجی لغوی است. از آنجا که هیچ‌گونه پیکره‌ای براساس محتوای وبی وجود ندارد؛ روش ارائه‌شده بر روی پیکره مقاله‌های یادگیری الکترونیکی پیاده‌سازی شده است. همچنین برای مقایسه کارایی روش ارائه شده با روش‌های قبلی دو روش پیشنهاد شده در ادبیات نیز بر روی پیکره مورد نظر پیاده‌سازی گشته است. نتایج نشان‌دهنده آن است که کارایی و دقت روش پیشنهادی بسیار بهتر از دو روش دیگر بوده و به کارگیری سنجش به صورت فازی تاثیر خوبی بر روی کارایی سنجش داشته است. روش ارائه شده در بیش از ۷۵٪ موارد، تحلیل درستی از شباهت دو سند ارائه می‌دهد و می‌توان از این روش برای جلوگیری از دستبرد ادبی و شناسایی اسناد مشابه استفاده کرد.

مراجع

[۱] . ساروخانی، لیلا و منتظر، غلامعلی. "طراحی و پیاده‌سازی سیستم هوشمند شناسایی رفتار مشکوک در بانکداری اینترنتی به کمک نظریه مجموعه‌های فازی"، فصلنامه فناوری اطلاعات و ارتباطات ایران، سال اول، شماره ۱و۲،صص ۹-۱۸،۱۳۸۷

[2] Wang, Yong, and Julia Hodges. "Document clustering with semantic analysis." System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on. Vol. 3. IEEE, 2006.

[3] Alzahrani, Salha M., Naomie Salim, and Ajith Abraham. "Understanding plagiarism linguistic patterns, textual features, and detection methods." Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 42.2 (2012): 133-149.

[4] Osman, Ahmed Hamza, Naomie Salim, Mohammed Salem Binwahlan, Rihab Alteeb, and Albaraa Abuobieda. "An improved plagiarism detection scheme based on semantic role labeling." Applied Soft Computing 12, no. 5 (2012): 1493-1502.

[5] Barrón-Cedeño, Alberto, and Paolo Rosso. "On automatic plagiarism detection based on n-grams comparison." In Advances in Information Retrieval, pp. 696-700. Springer Berlin Heidelberg, 2009.

[6] Kent, Chow Kok, and Naomie Salim. "Features based text similarity detection." arXiv preprint arXiv:1001.3487 (2010).

[7] Joy, Mike, and Michael Luck. "Plagiarism in programming assignments." Education, IEEE Transactions on 42.2 (1999): 129-133.

[8] Potthast, Martin, et al. "Cross-language plagiarism detection." Language Resources and Evaluation 45.1 (2011): 45-62.

[9] Baždarić, Ksenija. "Plagiarism detection–quality management tool for all scientific journals." Croatian medical journal 53.1 (2012): 1-3.

[10] Hariharan, Shanmugasundaram, et al. "Detecting plagiarism in text documents." Information Processing and Management. Springer Berlin Heidelberg, 2010. 497-500.

[11] Strapparava, Carlo, and Rada Mihalcea. "Semeval-2007 task 14: Affective text." Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics, 2007.

[12] Alzahrani, Salha, and Naomie Salim. "Fuzzy semantic-based string similarity for extrinsic plagiarism detection." Braschler and Harman (2010).

[13] Gupta, Rohit, et al. "UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment." SemEval 2014 (2014): 785.

[14] Huang,Y. -P. C.-C. Lu and T.-W. Chang , "An Intelligent Approach to Detecting the Bad Credit Card Accounts"," in 25th IASTED International Multi-Conference Artificial Intelligence and Applications, Innsbruck, Austria, pp. 1-6, 2007.

[15] Zadeh,L. A. "The concept of a linguistic variable and its application to approximate reasoning", Information Sciences, vol.8, pp. 199-249, 1975.

[16] Rutkowski,L. "Fuzzy Inference Systems", in Flexible Neuro-Fuzzy Systems, pp. 27-50, 2004

[17] Metzler, Donald, et al. "Similarity measures for tracking information flow." Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005.

[18] El-Alfy, El-Sayed M., et al. "Boosting paraphrase detection through textual similarity metrics with abductive networks." *Applied Soft Computing* 26 (2015): 444-453

زیرنویس‌ها

[1] Web
[2] Plagiarism
[3] Machine Translation
[4] Word Sense Disambiguation
[5] Summarizing
[6] Paraphrasing
[7] Text Clustering
[8] Similarity
[9] Vector Based
[10] Knowledge Based Database
[11] Uncertainty
[12] Vague
[13] Ontology
[14] Lexical Similarity
[15] N-Gram
[16] Skip-Gram
[17] Rosenfeld
[18] Precision
[19] Recall
[20] Confusion Matrix
[21] Positive Class
[22] Negative Class