طراحی سامانهٔ تشخیص دستبرد ادبی جمله‌بنیاد در متون فارسی به‌کمک هم‌جوشی گواه‌ها

حمید آهنگربهان و غلامعلی منتظر

دانشکدهٔ مهندسی صنایع و سیستم‌ها، دانشگاه تربیت مدرس، تهران، ایران

چکیده

امروزه، می‌توان با رونوشت‌برداری از منابع وسیع اینترنتی سند جدیدی را تهیه و به نام خود ارائه داد که مصداقی از دستبرد ادبی است. حفظ حقوق مالکیت مادی و معنوی نویسندگان ضرورت تشخیص اسناد اصلی و اسناد رونوشت‌برداری‌شده را ایجاب می‌کند. برای این منظور از الگوریتم‌های تشخیص دستبرد ادبی و سامانه‌هایی که مبتنی بر این الگوریتم‌ها طراحی شده‌اند استفاده می‌شود. عملکرد سامانه‌های تشخیص دستبرد ادبی که درواقع میزان شباهت بین دو متن را ارزیابی می‌کنند، بر اساس دو شاخص دقت یافته‌های مشابه و مدت زمان تشخیص سنجیده می‌شود. تاکنون سنجه‌های مختلفی برای ارزیابی مشابهت دو سند ارائه شده که کارایی آنها به محتوای متن و منابع مورد استفاده مانند واژه‌نامه برای مقایسهٔ بین واژه‌های دو سند محدود است. در این مقاله روشی ارائه شده است که از نظریهٔ گواه برای هم‌جوشی اطلاعات، به‌منظور ارزیابی تشابه دو سند و کشف دستبرد ادبی با توجه به کیفی و ناکامل‌بودن عوامل اثرگذار بر سنجش شباهت بین دو متن، استفاده می‌کند. روش پیشنهادی در گام نخست جمله‌های موجود در سند را به دو بخش عمومی و تخصصی تقسیم کرده و سپس با استفاده از سنجه‌های متفاوت و استفاده از «هستان‌نگار تخصصی» امتیاز تشابه برای هر بخش را محاسبه و درنهایت در دو سطح، میزان شباهت بین دو سند را استنتاج می‌کند؛ در سطح نخست نتایج سنجه‌های شباهت‌سنجی به‌عنوان گواه (با باور پایهٔ مشخص) با قاعدهٔ دمپستر-شفر با هم ترکیب شده و به‌عنوان گواهی جدید به سطح دوم منتقل می‌شوند. در سطح دوم نتیجهٔ سطح نخست و گواه جدید از طریق قاعدهٔ میانگین‌گیری ترکیب شده و توابع باور و مقبولیت نهایی محاسبه و شباهت بین دو جمله (سند) ارزیابی می‌شود. سامانهٔ مذکور بر روی پیکرهٔ فارسی که با استفاده از مقاله‌های کنفرانس سالانهٔ یادگیری و آموزش الکترونیکی تولید شده، مورد ارزیابی قرار گرفته است که با دقت بیش از ۸۶٪ امکان شناسایی اسناد مشابه را دارد.

واژگان کلیدی: دستبرد ادبی، هم‌جوشی داده، سنجه‌های شباهت‌سنجی، نظریهٔ گواه، شباهت‌سنجی معنایی.

۱- مقدّمه

با فراگیرشدن شبکهٔ وب و حجم زیاد اطلاعاتی که این شبکه در اختیار قرار می‌دهد، چالش‌های بسیاری در حوزهٔ پردازش زبان‌های طبیعی به‌وجود آمده است. یکی از روش‌های مؤثر برای کاهش و سازماندهی این حجم زیاد داده‌ها، شناسایی اسناد مشابه و دسته‌بندی آنهاست(ونگ و هوجز، ۲۰۰۶). با یافتن اسناد مشابه امکان هر نوع استفادهٔ غیرقانونی، که دستبرد ادبی نمونهٔ بارز آن است، محدود خواهد شد.

در تشخیص اسناد جعلی از اسناد اصل، انتخاب درست ویژگی‌ها و روش‌های تشابه‌یابی متن جنبهٔ کلیدی دارد(هرمان وزاکا،۲۰۰۶). به همین دلیل نیاز است که فرآیند به‌کارگیری اسناد توسط کاربران به‌دقّت بررسی شود. در حال حاضر، با توسعهٔ شبکه‌های اطلاعاتی و به‌ویژه وب، کاربران به‌راحتی می‌توانند سند جدیدی را با رونوشت‌برداری از منابع موجود به نام خود تهیه کنند که طبعاً مصداقی از دستبرد ادبی محسوب می‌شود. البته گاهی کاربر برای رد گم‌کردن از واژه‌های مترادف یا تغییر ساختار جمله‌ها استفاده می‌کنند که در این صورت تشخیص این نوع دستبرد به‌صورت خودکار مشکل است و سامانه‌هایی همانند «کُپس»[1] (عثمان و همکاران،۲۰۱۳) ، «اِسکم»[2] (رام و همکاران،۲۰۱۴) و «چک»[3] (کایا و همکاران،۲۰۱۴)، قابلیت

[1] COPS
[2] SCAM
[3] CHECK

شناسایی آن را ندارند. اشکال اصلی این سامانه‌ها، کُندی، وابستگی به ساختار لغوی و عدم توجه به ساختار معنایی جمله‌ها و واژه‌هاست(عثمان و همکاران،۲۰۱۳؛باو و همکاران ۲۰۰۳). بنابراین تشخیص متون هم‌معنانویسی[1] شده، برای آنها بسیار دشوار خواهد بود. به همین دلیل چالش اصلی در این زمینه یافتن الگوریتمی مناسب برای بهبود دقت یافته‌های مشابه در مقایسهٔ دو متن است که ساختار لغوی و معنایی جمله را همزمان در نظر داشته باشد. شایان توجه است که اگرچه مدت زمان سنجش شباهت دو متن نیز می‌تواند از شاخص‌های ارزیابی سامانه‌های تشخیص دستبرد ادبی باشد، اما از آنجا که مسئله دستبرد ادبی به‌طورمعمول به‌صورت برون‌خط انجام می‌شود، دقت تشخیص بیش از مدت زمان آن مورد توجه است. سامانه‌های دستبرد ادبی از روش‌های اندازه‌گیری شباهت بین اجزای دو سند، چه در سطح لغوی و چه در سطح معنایی، استفاده می‌کنند از طرفی در دنیای واقعی، هر سند مجموعه‌ای از محتواهای متنوع و مختلف است و باید روشی را یافت که اطلاعات محتوای متن را به‌خوبی نشان دهد و سنجهٔ شباهت‌سنجی و منبع اطلاعاتی مورد نیاز در همان زمینه و محتوا را در روش‌های شباهت‌سنجی به کار برد. از آنجا که محتوای سند به چندین بخش (از جمله تخصصی و عمومی) تقسیم می‌شود، به همین دلیل لازم است، سامانه، از منابع اطلاعاتی مختلف برای تصمیم‌گیری دربارهٔ شباهت بین دو سند استفاده کند تا دیدی جامع‌تر و معقول‌تر نسبت به مسئله داشته و درنهایت اطلاعات حاصل از هر بخش را با هم ترکیب کند. هم‌جوشی داده‌ها راه‌حلی برای یک‌پارچه‌سازی اطلاعات حاصل از منابع مختلف است و می‌توان آن را به‌صورت «فرآیند چندسطحی و چندوجهی برای تشخیص، تجمیع، اصلاح، برآورد و ترکیب خودکار داده و اطلاعات از منابع مختلف» تعریف کرد(وایت، ۱۹۹۱؛ دونگ و همکاران،۲۰۰۹). این روش در مسائل بسیاری، مانند پردازش تصویر، پردازش سیگنال و استنتاج مورد استفاده قرار گرفته است(لنز و همکاران،۲۰۱۲).

در این مقاله پس از بررسی سامانه‌های تشخیص دستبرد ادبی موجود ، مزایا و معایب آنها مطرح می‌شود و سپس روشی مبتنی بر هم‌جوشی داده‌ها[2] و به‌کمک نظریهٔ گواه[3] و در قالب سامانهٔ تشخیص دستبرد ادبی برای شناسایی اسناد مشابه فارسی ارائه خواهد شد.

[1] Paraphrase
[2] Data Fusion
[3] Evidence Theory

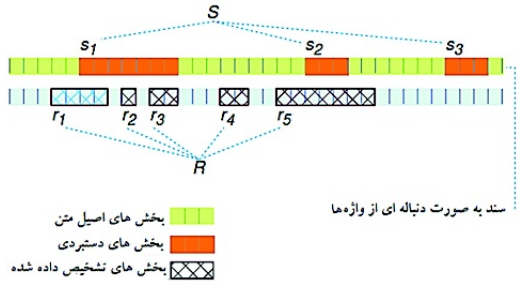
ادامهٔ این مقاله بدین صورت تنظیم شده است: در بخش۲ مسئلهٔ دستبرد ادبی تعریف می‌شود. در بخش ۳ مروری بر کارهای پیشین در زمینه سامانه‌های شناسایی دستبرد ادبی، روش‌های شباهت‌سنجی و در بخش ۴ مفاهیم اصلی نظریهٔ گواه ارائه و روش‌های ترکیب گواه در این نظریه بحث می‌شود. در بخش ۵ معماری پیشنهادی برای سامانهٔ شباهت‌سنجی بیان می‌شود؛ پیاده‌سازی این سامانه و نتایج عملی حاصل از کاربرد آن در بخش ۶ بیان خواهد شد و درنهایت بخش ۷ به نتیجه‌گیری می‌پردازد.

۲- تعریف دستبرد ادبی

دستبرد ادبی[4] در ساده‌ترین شکل به‌صورت «رونوشت‌برداری از اسناد بدون تأیید مرجع اصلی» تعریف می شود که یکی از شایع‌ترین موارد آن، استفاده از مطالب دیگران در نوشته‌ای به نام خود است(جوی،۱۹۹۹).

فرآیند تشخیص دستبرد ادبی را به کمک نمادهای ریاضی به صورت زیر می‌توان تعریف کرد(استین و همکاران،۲۰۰۹): مجموعهٔ مرجع D و مجموعه اسناد مشکوک D_q مفروض است، فرآیند تشخیص دستبرد ادبی عبارت است از یافتن بخش‌های مشکوک s_q از D_q که مشابه بخش‌های s_x در مجموعهٔ مرجع D است.

شکل (۱) این فرآیند را نشان می‌دهد. در این فرآیند سعی بر این است تا بخش‌های تشخیص‌داده‌شده بیش‌ترین تطابق را با بخش های دستبردی داشته باشد. به‌عنوان نمونه در شکل ۱، بخش S_3 دستبردی تشخیص داده نشده است:



(شکل- ۱): فرآیند تشخیص دستبرد ادبی

انواع مختلف دستبرد ادبی عبارتند از(موشکه و گیپ،۲۰۱۳):

۱- دستبرد تحت اللفظی[5]: این نوع دستبرد عبارت است از رونوشت برداری متن بدون تغییر یا با تغییرات کم.

[4] Plagiarism
[5] Literal plagiarism

«رونوشت برداری و چسباندن»[1] از متون منابع دیگر، اصلی‌ترین شکل این نوع دستبرد است(وبر- وولف،۲۰۱۱).

۲- دستبرد پنهانی[2]: در این نوع دستبرد از شیوه‌هایی برای پنهان‌کردن رونوشت برداری متن استفاده می شود. سه نوع دستبرد پنهانی وجود دارد:

۲-۱- قطعه‌کردن و چسباندن[3]: در این شیوه از رونوشت‌برداری و تجمیع تکه متن‌ها با تغییرات جزئی (مانند تغییر ترتیب واژه‌ها، جایگزینی آنها با مترادف‌شان و یا حذف و اضافه کردن واژه‌ها) استفاده می‌شود(وبر- وولف،۲۰۱۱).

۲-۲- هم‌معنانویسی[4]: عبارت است از بازنویسی خودآگاهانهٔ ایده‌های دیگران به‌صورت لغوی و نگارش آن بدون اعلام منبع اصلی(لنکستر،۲۰۰۳).

۲-۳- پنهان‌کاری فنی[5]: در این روش متونی ساخته می‌شود که برای ماشین قابل تشخیص نباشد. در این شیوه از روش‌هایی مانند جایگزینی نویسه‌ها با نمادهای گرافیکی و الفبای خارجی یا اضافه‌کردن تصادفی کلمات با قلم سفید استفاده می‌شود(هیثر،۲۰۱۰).

۳- دستبرد ترجمه ای[6]: این نوع دستبرد ادبی عبارت است از تبدیل خودکار یا دستی و آگاهانهٔ متن از یک زبان به زبان دیگر بدون بیان منبع اصلی آن(وبر- وولف،۲۰۱۱).

۴- دستبرد ایده[7]: شامل استفاده از مفاهیم و ایده‌های دیگران مانند تصاحب دیدگاه‌های پژوهشی بدون اطلاع‌رسانی دربارهٔ منبع آن(مورر و همکاران،۲۰۰۶).

۵- خوددستبردی[8]: به معنای استفاده کامل یا قسمتی از نوشته‌های خود بدون ارجاع به اصل(براتگ و محمود،۲۰۰۹).

۳- روش‌های شناسایی دستبرد ادبی

در این بخش کارهای پژوهش‌گران در زمینهٔ دستبرد ادبی و شباهت‌سنجی متون بررسی می‌شود. دو سند را مشابه گوییم اگر میزان شباهت واژه‌های آنها چه از نظر لغوی و چه معنایی از آستانه‌ای معلوم بیشتر باشد(گراوند و همکاران،۲۰۱۴). این آستانه با توجه به نوع رویکرد سنجش (معنایی و یا لغوی) و نیز دقت در اندازه‌گیری شباهت مشخص می‌شود. به‌طورمعمول مقدار این آستانه در سامانه‌های تشخیص تشابه متغیر در نظر گرفته شده و بر اساس خواست کاربر تنظیم می‌شود. یکی از سامانه‌های طراحی‌شده برای شناسایی شباهت بین متون، سامانهٔ «کُپس» است که در آن توابع رمزگذاری[9] برای تولید اثرانگشت[10] به کار گرفته شده و مقادیر رمزنگاری دو سند برای تشخیص شباهت استفاده می‌شود(برین و همکاران،۱۹۹۵). سامانهٔ «اِسکم» نیز از روش‌های بازیابی اطلاعات مبتنی بر واژه برای تشخیص رونوشت‌برداری استفاده می‌کند که تنها قابلیت شناسایی اسناد کوتاه را دارد (شیواکومار و همکاران،۱۹۹۵). سامانهٔ «چک» با هدف کاهش پیچیدگی، ابتدا تعداد کمی از ویژگی‌های دو سند را مورد بررسی قرار می‌دهد و در صورت برقرار بودن شرایط نخستین الگوریتم روی کل دو سند اجرا می‌شود(لنز و همکاران،۱۹۹۷). در سامانه «تشخیص رونوشت‌برداری از کالاهای دیجیتال»[11] نیز از روش‌های مبتنی بر بردار، برای تشخیص شباهت استفاده شده است(باو و همکاران،۲۰۰۱). برای بررسی جمله به جملهٔ دو سند نیز سامانهٔ «اسپلت»[12] پیشنهاد شده که با محاسبهٔ تعداد تکرار واژه‌ها این کار را انجام می‌دهد(کالبرگ و همکاران،۲۰۰۳). روش‌های دیگری از جمله، «نشان‌گذاری دیجیتالی»[13] که اطلاعات اضافه‌ای را به سند ضمیمه می‌کند نیز برای استفاده در این سامانه‌ها پیشنهاد شده است(هیاری و همکاران،۲۰۰۵؛زینی و همکاران، ۲۰۰۶). روش‌های خوشه‌بندی اسناد نیز یکی از روش‌های بازیابی اطلاعات است که بازیابی داده را با کاهش زمان جستجو در جایابی اسناد، خلاصه‌سازی متون و کاهش زمان مقایسه در تشخیص دستبرد ادبی بهبود می‌بخشد. در برخی پژوهش‌ها نیز از واژه‌های کلیدی برای یافتن خوشه‌های مشابه بین اسناد استفاده شده است(برین و همکاران،۱۹۹۵).

پژوهش‌های بسیاری نیز برای تشخیص دستبرد ادبی بر ویژگی‌های ساختاری متن مانند سرسندها[14]، بخش‌ها، پاراگراف‌ها و مراجع تمرکز دارند. چاو و رحمان از روش‌های نمایش متن به‌صورت ساختار درختی استفاده کرده و به کمک شبکهٔ نگاشت خودسازماندهٔ چندلایه[15] آن را غنی و

[1] Copy and paste
[2] Disguised plagiarism
[3] Shake and paste
[4] Paraphrasing
[5] Technical disguise
[6] Translated plagiarism
[7] Idea plagiarism
[8] Self-plagiarism

[9] Hashing
[10] Fingerprint
[11] Copying Detection System of Digital Goods
[12] SPLaT
[13] Digital Watermarking
[14] Header
[15] Multi-Layer Self-Organizing Maps Model

برای تشخیص دستبرد ادبی به کار برده‌اند(چاو و رحمان،۲۰۰۹). در روش پیشنهادی اقبال و همکاران، از رویکرد مرسوم مبتنی بر بردار و همچنین روش‌های مبتنی بر گراف برای تشخیص پاراگراف‌های مشکوک استفاده شده است(اقبال و همکاران،۲۰۱۲). گراوند و همکاران نیز روشی مبتنی بر فیلترهای بلوم[1] را برای بررسی دستبرد ادبی در اطلاعات شخصی و محرمانه برای حفظ اطلاعات افراد پیشنهاد کرده‌اند(گراوند،۲۰۱۴).

اصلی‌ترین بخش هر سامانهٔ شناسایی دستبرد ادبی، الگوریتم شباهت‌سنجی آنهاست ازاین‌رو در این بخش مزایا و معایب هر رویکرد برای درک بهتر مسئله تشخیص دستبرد ادبی، بیان می‌شود:

تشابه متن را می‌توان در دو حوزهٔ کلی « تشابه لغوی[2] » و «تشابه معنایی[3] » بررسی کرد. تشابه‌یابی لغوی، به ظاهر واژه‌ها توجه می‌کند و برای سنجش آن در ابتدا واژه‌ها ریشه‌یابی‌شده و سپس از طرق مختلفی تشابه اندازه‌گیری می‌شود(متزلر و همکاران،۲۰۰۷). «تشابه معنایی» نیز مفهومی است که تشابه بین مجموعه‌ای از اسناد یا واژه‌ها را از طریق معنای آنها مشخص می‌کند(میهالیشا و همکاران،۲۰۰۶). رویکردها و روش‌های تشابه‌یابی و همچنین مقایسه این روش‌ها با شاخص‌هایی که در پیاده‌سازی نرم‌افزاری آنها و همچنین تشخیص دستبرد ادبی مهم است، در جدول (۱) نمایش داده شده است. ضعف اصلی روش‌های تشابه‌یابی لغوی این است که نمی‌توانند عبارات متفاوت را، که از نظر معنایی با هم مرتبط هستند، شناسایی کنند؛ در تشابه‌یابی معنایی نیز چالش اصلی پایگاه دانش و پیکرهٔ غنی و جامعِ مورد نیاز است؛ به‌عنوان نمونه سنجه‌های مبتنی بر پایگاه دانش، فرآیند رمزنگاری خاصی را برای روابط معنایی دارند، در مقابل، سنجه‌های مبتنی بر پیکره، چون از داده‌های خام و بی‌ساختار استفاده می‌کنند به‌صورت گسترده قابل استفاده هستند؛ ولی متأسفانه این ویژگی‌ها برای این دو دسته سنجه دارای هزینه است و محدودیت‌های مشترکی را برای آنها به‌وجود می‌آورد که عبارتند از: هزینهٔ محاسباتی و نیازمندی به حافظهٔ زیاد، بی‌میلی برای استفاده در چند زبان، کمبود پایگاه دانش با کیفیت بالا، برآورد ضعیف از ارتباط معنایی و عدم توانایی برای استفاده در حوزهٔ خاص (محتوا محور)(محمد،۲۰۰۸). ضعف‌ها و محدودیت‌های سنجه‌های تشابه لغوی و معنایی باعث می‌شود که این سنجه‌ها میزان شباهت متفاوتی را برای زوج‌واژه‌ها در نظر بگیرند، به همین دلیل اگر سندی شامل واژه‌های یک حوزهٔ خاص در اختیار داشته باشیم، این سنجه‌ها به‌تنهایی کارایی مناسبی نخواهند داشت؛ لذا در این‌گونه متون استفاده از سنجه‌ای ترکیبی برای رفع این معایب احساس می‌شود. جدول (۲) بیان‌گر دسته‌بندی سامانه‌های پیشنهادی برای تشخیص دستبرد ادبی است. این جدول نشان می‌دهد تاکنون تمرکز پژوهش‌گران به استفاده از روش‌های شباهت‌سنجی لغوی بوده و کمتر از شباهت‌سنجی معنایی استفاده کرده‌اند. سنجه‌های شباهت‌سنجی معنایی نیز به‌شدت وابسته به ساختار و کیفیت پایگاه دانش است. به همین دلیل در این مقاله به‌دنبال طراحی سامانه‌ای برای تشخیص دستبرد ادبی هستیم که بتواند مشکلات موجود در روش‌های شباهت‌سنجی لغوی و معنایی را پوشش داده و از مزایای هر دو روش نیز استفاده کند.

۴- نظریهٔ گواه دمپستر- شفر

شناخت عوامل اثرگذار بر مسائل مهندسی در دنیای امروزی، اصلی‌ترین عنصر حل آن خواهد بود. در دنیای واقعی هر داده و اطلاعاتی به‌طورمعمول با نوعی نقصان[4] روبه‌رو است که برای حل آنها باید از روش‌های متناسب با نوع نقص آن استفاده کرد. در حال حاضر نظریه‌های زیادی برای نمایش نقص داده مانند نظریهٔ احتمال[5]، نظریهٔ مجموعهٔ فازی[6]، نظریهٔ امکان[7]، نظریهٔ مجموعه نادقیق[8] و نظریهٔ گواه دمپسترشفر مورد استفاده قرار می‌گیرد. بسیاری از این نظریه‌ها توانایی بازنمایی دادهٔ ناقص را تنها در منظری خاص دارند. برای مثال توزیع احتمال، قابلیت بیان عدم اطمینان در داده و نظریهٔ مجموعه فازی توانایی بازنمایی داده مبهم با استفاده از تابع عضویت فازی را دارد. این نظریهٔ تنها توانایی حل مسائل باابهام را به‌وسیله متغیرهای زبانی و هم‌جوشی از طریق قواعد فازی دارد درحالی‌که نظریهٔ گواه، توانایی هم‌جوشی دادهٔ نامطمئن و نامعلوم را دارد و در حل انواع دیگر داده ناقص ناتوان و برای هم‌جوشی داده با تعارض بالا نیز ناکارآمد است(خالقی و همکاران،۲۰۱۳).

[1] Bloom filters
[2] Text or Lexical Similarity
[3] Semantic Similarity
[4] Imperfection
[5] Probabilty Theory
[6] Fuzzy Sets Theory
[7] Possibilty Theory
[8] Rough Set Theory

(جدول- ۱): رویکردها و روش‌های شباهت‌سنجی(محمد،۲۰۰۸)

نوع تشابه	رویکرد	روش	دقت اندازه‌گیری	حافظه مورد نیاز	زمان پردازش	پیچیدگی محاسبات	پایداری در برابر تغییرات ویرایشی
لغوی	**هندسی** (بارون و همکاران،۲۰۰۹)	هم‌پوشانی واژه‌ای، سنجه‌های فراوانی واژه-وارون فراوانی سند، فاصله ویرایشی، شباهت کسینوسی	متوسط	زیاد	زیاد	کم	کم
	ویژگی‌محور	مدلهای اثرانگشتی، اثرانگشت غربالی(کونت کنت و همکاران،۲۰۱۰)، الگوریتم‌های مبتنی بر ترتیب واژه(ژانگ و همکاران،۲۰۱۱)	خوب	زیاد	زیاد	زیاد	متوسط
	مبتنی بر احتمال (بارون و همکاران،۲۰۰۹)	ترجمه ماشینی، Okapi BM25، فاصله کالبک-لیبلر	متوسط	زیاد	زیاد	زیاد	خوب
	شبکه استنتاجی (آدیس،۲۰۱۲)	-	بسته به داده آموزشی	زیاد	زیاد	متوسط	متوسط
	یادگیرندهٔ فاصله (بیلنکو و مونی، ۲۰۰۵)	فاصله ویرایشی رشته یادگیری پذیر با شکاف وابستگی، فضای برداری یادگیری‌پذیر	بسته به داده آموزشی	زیاد	متوسط	متوسط	متوسط
معنایی	**مبتنی بر پایگاه دانش** (میهالیشا و همکاران، ۲۰۰۶)	وو و پالمر، رزنیک،هیرست و اس تی آنگ،لی لی کوک و چادرو،لسک	بسته به پایگاه دانش	زیاد	زیاد	زیاد	زیاد
	مبتنی بر پیکره (محمد،۲۰۰۸)	اطلاعات متقابل نقطه‌ای،تحلیل معنایی پنهان،تحلیل معنایی پنهان احتمالی،تخصیص دیرکله پنهان،مدل ابرفضای قیاس به زبان،نمایه سازی تصادفی	بسته به پایگاه پیکره	زیاد	زیاد	زیاد	کم
	روش‌های هوشمند	شبکه‌های عصبی(یح و همکاران،۲۰۱۱) ، مجموعه فازی(گوپتا و همکاران،۲۰۱۴)، نظریه گواه(لی و همکاران،۲۰۰۶)	خوب	زیاد	زیاد	زیاد	زیاد

(جدول- ۲): دسته‌بندی سامانه‌های تشخیص دستبرد ادبی

نام	رویکرد شباهت‌سنجی	نوع تشابه	روش مقایسه
منبر (عثمان و همکاران،۲۰۱۳)	هندسی	لغوی	واژه به واژه
کپس (عثمان و همکاران،۲۰۱۳)	هندسی	لغوی	اثرانگشت
اسکم (رام و همکاران،۲۰۱۴)	هندسی	لغوی	واژه به واژه(در سطح جمله)
چک (کایا و همکاران،۲۰۱۴)	ویژگی‌محور	لغوی	ویژگی‌های جمله
رونوشت‌برداری از کالاهای دیجیتال (باو و همکاران،۲۰۰۱)	هندسی	لغوی	مبتنی بر بردار
اسپلت (کولبرگ و همکاران،۲۰۰۳)	هندسی	لغوی	تعداد تکرار واژه‌ها
نشان‌گذاری شفاف دیجیتال (زینی و همکاران،۲۰۱۳)	ویژگی‌محور	لغوی	اضافه‌کردن اطلاعات به سند
زینی (زینی و همکاران،۲۰۱۳)	ویژگی‌محور	لغوی/معنایی	خوشه‌بندی
چاو و رحمان (چاو و رحمان،۲۰۰۹)	ویژگی‌محور	لغوی/معنایی	شبکه‌های عصبی
اقبال (اقبال و همکاران،۲۰۱۲)	هندسی	لغوی	مبتنی بر بردار و گراف
گراوند (گراوند و همکاران،۲۰۱۴)	هندسی	لغوی	استفاده از فیلترهای بلوم

در مسائل دنیای واقعی بخشی از نقص اطلاعات و داده ناشی از شناخت ناقص فضای حالت است به‌طوری‌که عوامل اثرگذار بر مسئله و همچنین فضای حالت آن به‌راحتی قابل شناسایی نیست. در این موارد، عدم اطمینانی نسبت به فضای حالت مسئله به‌وجود می‌آید که با روش‌های احتمالی قابل بررسی نیست. این عدم اطمینان نوعی عدم قطعیت به مسئله وارد می‌کند. به‌عنوان مثال، آیا با فرض دانستن اینکه به‌احتمال ۱۰٪ در حیاط دانشگاه نفت وجود ندارد، می‌توان به این اطمینان رسید که به‌احتمال ۹۰٪ در حیاط دانشگاه نفت وجود دارد؟ آرتور دمپستر[1] نخستین‌بار در دهه ۱۹۶۰ به این موضوع اشاره کرد و تلاش کرد تا به‌جای استفاده از یک عدد احتمالیِ منفرد برای نمایش عدم قطعیت؛ آن را به‌صورت بازه‌ای از احتمالات مدل کند تا بتواند شناخت ناکاملی را از فضای حالت مسئله بیان نماید(دمپستر،۱۹۶۷). شفر[2] ضمن اصلاح و گسترش کار دمپستر، کتابی را به نام «نظریهٔ ریاضی گواه»[3] در سال ۱۹۶۷ منتشر کرد(شفر،۱۹۷۵). کار گسترده‌تر شفر «استدلال مبتنی بر گواه»[4] نام داشت که به تفسیر اطلاعات غیرقطعی، نادقیق و برخی ناصحیح می‌پرداخت(گریور،۱۹۸۱).

اجزای اصلی این نظریهٔ گواه شامل: چارچوب تشخیص[5]، توابع انتساب باور پایه[6]، باور[7] و مقبولیت[8] است که در ادامه توضیح داده می‌شود:

۴-۱- چارچوب تشخیص

«چارچوب تشخیص»، مجموعه‌ای از گزاره‌های کامل و منحصر به فرد (همانند فضای حالت در نظریهٔ احتمال در بررسی مسئله است) که چارچوبی را برای پوشش همهٔ حالت‌های مسئله ارائه می‌کند. مجموعه چارچوب تشخیص را با نماد Θ و با رابطه ۱ نمایش می‌دهند(گوآن و بل،۱۹۹۱):

$$\Theta = \{\theta_1, \theta_2, ..., \theta_N\} \quad (1)$$

که θ_i حالت‌های ممکن رخداد در فضای حالت را نشان می‌دهد و منظور از «تشخیص» این است که امکان تشخیص یک پاسخ صحیح از میان همهٔ پاسخ‌های ممکن وجود دارد. اگر پاسخ در چارچوب نباشد، نشان‌دهندهٔ آن است که شناخت کاملی از فضای حالت مسئله وجود نداشته و باید چارچوب توسعه پیدا کند تا با دانش اضافه‌شدهٔ عناصر $\theta_{N+1}, \theta_{N+2}$ و ... تطبیق یابد(گیارانتانو،۲۰۰۵).

۴-۲- تابع انتساب باور پایه

مقداردهی در نظریهٔ گواه با «تابع انتساب باور پایه» انجام می‌شود. این شاخص را می‌توان با استفاده از نگاشتی مانند m برای بیان باور خود دربارهٔ یک گزاره، به شکل عددی در بازهٔ $[0,1]$ ارائه داد(خطیبی و منتظر،۲۰۱۰):

$$m : 2^\theta \rightarrow [0,1] \quad (2)$$

این تابع دارای خواص زیر است:

الف- برای هر یک از گزاره‌ها، باور می‌تواند مقداری بزرگ‌تر یا مساوی صفر داشته باشد؛ یعنی:

$$\forall A \in 2^\theta \rightarrow m(A) \geq 0 \quad (3)$$

ب- هیچ درجه‌ای از باور برای گزارهٔ تهی در نظر گرفته نمی‌شود؛ یعنی:

$$m(\phi) = 0 \quad (4)$$

ج- مجموع کل باور برابر با واحد است؛ یعنی:

$$\sum_{A \in 2^\theta} m(A) = 1 \quad (5)$$

۴-۳- تابع باور و تابع مقبولیت

«تابع باور» نشان‌دهندهٔ درجهٔ باور کامل به گزاره‌ای است که نسبت به آن اطمینان وجود دارد. در این تابع، تنها گزاره‌هایی وارد می‌شوند که به شکل کامل مؤید گزارهٔ مدّنظر هستند. این تابع به‌صورت رابطه (۶) تعریف می‌شود(بای،۲۰۰۴):

$$Bel(A) = \sum_{C_i \in 2^\theta, C_i \subset A} m(C_i) \quad (6)$$

که $m(C_i)$ بخشی از باور کامل منتسب به گزارهٔ A است.

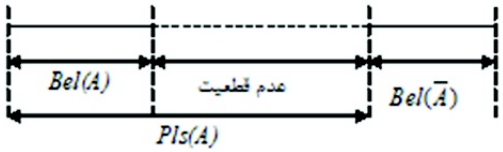
«تابع مقبولیت» نیز بیان‌گر درجهٔ مقبولیت یک گزاره است. درواقع درجهٔ مقبولیتِ یک گزاره، به معنای حد بیشینهٔ وقوع آن است. چون هر گزاره‌ای که با گزارهٔ مدّنظر اشتراک دارد، حدّاقل بخشی از آن گزاره را پوشش می‌دهد و بدین‌ترتیب درجهٔ باور گزارهٔ دارای اشتراک بر امکان‌پذیری و مقبولیت گزارهٔ مدّنظر دلالت می‌کند. این تابع به‌صورت رابطهٔ (۷) تعریف می‌شود(گوآن و همکاران،۱۹۹۱):

$$Pls(A) = \sum_{C_i \cap A \neq \phi} m(C_i) \quad (7)$$

---

[1] Arthur Dempster
[2] Shafer
[3] A mathematical theory of evidence
[4] Evidential reasoning
[5] Frames of discernment
[6] Basic probability assignment function
[7] Belief function
[8] Plausibility function

درصورت فقدان اطلاعات و دانش کافی دربارهٔ مسئله، بسیار معقول است که به جای ارایهٔ یک مقدار معیّن برای متغیر نامطمئن (همانند آنچه که در نظریهٔ احتمال انجام می‌شود) بازه‌ای برای محدودهٔ عدم‌قطعیت ارائه شود. با توجه به تعریف تابع باور، این تابع حداقل درجهٔ گواه کامل گزاره را که دربارهٔ آن اطمینان وجود دارد، ارائه می‌کند؛ امّا تابع مقبولیت، حداکثر درجهٔ باور بالقوهٔ گزاره را بیان می‌کند. بدین ترتیب، مقدار متغیر عدم‌قطعی (درجهٔ باور) بین دو کران پایین و بالا، که به‌ترتیب تابع باور و تابع مقبولیت است، محصور می‌شود. همان طور که در شکل ۲ ملاحظه می‌شود، درجهٔ باور گزارهٔ فرضی A با بازهٔ [Bel(A),Pls(A)] ارائه می‌شود. با توجه به اینکه مقادیر توابع باور و مقبولیت در بازهٔ [۰,۱] هستند، درجهٔ باور گزاره نیز در بازهٔ [۰,۱] جای دارد(خطیبی و منتظر،۲۰۱۰).



(شکل- ۲): توابع باور و مقبولیت یک گزاره

۴-۴- ترکیب گواه

اگر چند تابع انتساب گواه پایه برای چارچوب تشخیص مسئله ارائه شود، باید آنها را با هم ترکیب کنیم، تا تابع انتساب جدیدی حاصل شود که حاوی معتبرترین و کامل‌ترین اطلاعات باشد(گوآن و همکاران،۱۹۹۱). به‌لحاظ ریاضی، دو گواه m_1 و m_2 را که از طریق دو منبع اطلاعاتی مستقل حاصل آمده، می‌توان با استفاده از قاعدهٔ ترکیب گواه ترکیب کرد و به تابع انتساب جدیدی دست یافت؛ این ترکیب با رابطهٔ (۸) تعریف می‌شود:

$$m(A)=\frac{\sum_{c_i\cap c_j=A} m_1(C_i)m_2(C_j)}{1-\sum_{C_i\cap C_j=\Phi} m_1(C_i)m_2(C_j)},\qquad A\neq\phi \qquad (8)$$

در این قاعده، C_i و C_j هر یک به گزاره‌هایی از یک منبع مستقل اشاره می‌کنند. روش میانگین‌گیری[1] صحّت اطلاعات ساختارهای گواه را یکسان فرض می‌کند و از تناقض موجود بین ساختارهای گواه چشم‌پوشی می‌کند. قاعدهٔ ترکیب میانگین‌گیری به‌صورت رابطهٔ (۹) است(بای،۲۰۰۴):

$$m_{1..n}=\frac{1}{n}\sum_{i=1}^{n} w_i m_i \qquad (9)$$

در این رابطه m_i تابع انتساب گواه پایهٔ گزاره و w_i وزن این انتساب (بر اساس اعتبار اطلاعات موجود) است.

۵- معماری‌سامانهٔ تشخیص دستبرد ادبی

شکل (۳) معماری سامانهٔ پیشنهادی را نشان می‌دهد. در این معماری مراحل زیر به‌ترتیب بر روی دو سند مشکوک و مرجع انجام خواهد شد:

۱- قطعه‌بندی و ثبت سند در پایگاه داده

۱-۱- قطعه‌بندی‌کردن جمله‌های اسناد براساس فهرستی از علائم معیار همانند نقطه‌گذاری و نقطه‌ویرگول

۱-۲- ثبت جمله‌ها و ترتیب آنها در پایگاه داده

۱-۳- قطعه‌بندی جمله‌ها براساس معیار «فاصله» بین واژه‌ها

۱-۴- ثبت واژه‌ها به‌همراه ترتیب آنها برای ساخت چندکلمه‌ها در پایگاه داده

۱-۵- تولید چندکلمه‌های واژه برای تولید عبارت چندواژه‌ای

۲- پیش‌پردازش سند

۲-۱- پیش‌پردازش واژه‌ها و عبارت چندواژه‌ای و حذف ایست‌واژه‌ها (مانند «از» و «به»)

۲-۲- ریشه‌یابی واژه‌ها و عبارت چندواژه‌ای (به‌طورمثال ریشه واژه «خور» «می‌خورند» است.)

۳- تحلیل سند

۳-۱- ساخت ماتریس تعداد تکرار واژه و اهمیت واژه (TF-IDF) از روی واژه‌ها و عبارت چندواژه‌ای ساخته شده

۳-۲- یافتن واژه‌ها و عبارت چندواژه ای اصلی براساس تعداد وقوع

۳-۳- جداسازی واژه‌های تخصصی براساس هستان‌نگار موضوعی

۳-۴- تشکیل ماتریس جمله‌های تخصصی و عمومی

۴- محاسبه شباهت بین اسناد

۴-۱- اندازه‌گیری میزان شباهت زوج‌جمله‌ها بر اساس ماتریس جمله‌های عمومی دو سند با استفاده از واژه‌نامه

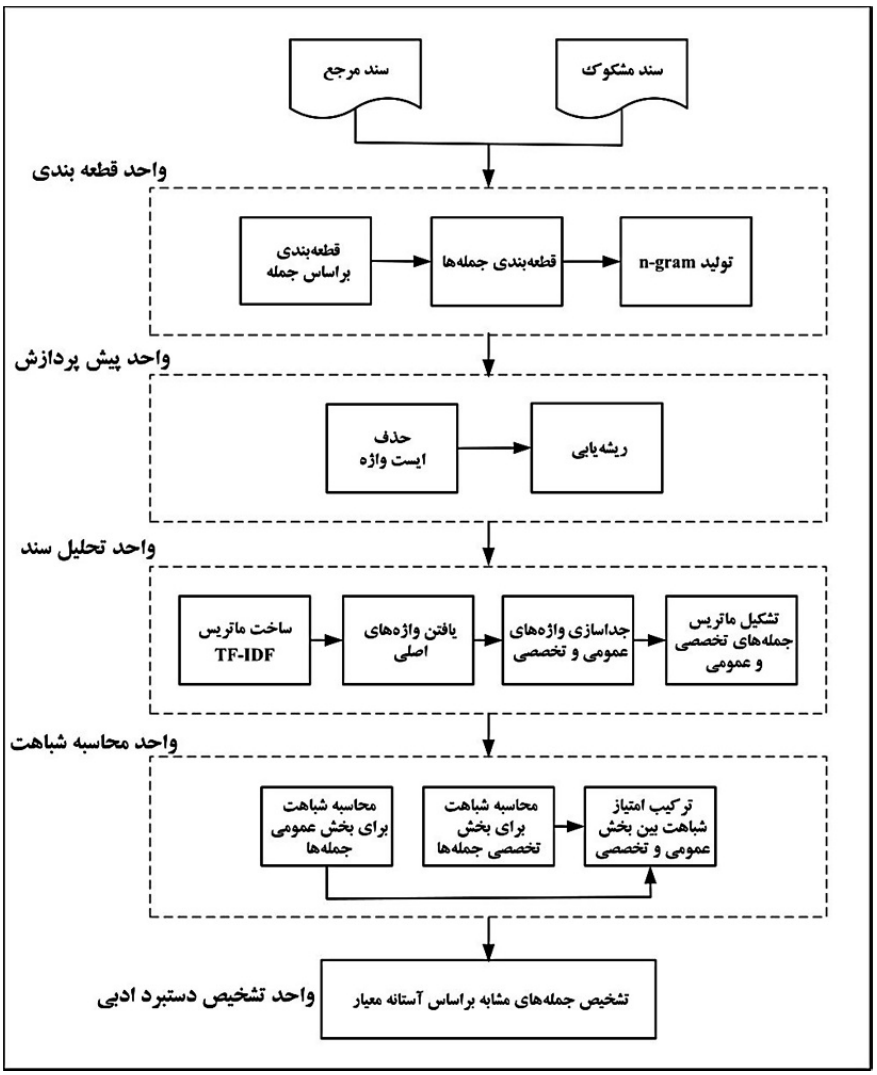
[1] Averaging Method

۴-۲- اندازه‌گیری میزان شباهت زوج جمله‌ها براساس ماتریس جمله‌های تخصصی دو سند با استفاده از هستان‌نگار موضوعی

۴-۳- ترکیب امتیاز شباهت عمومی و تخصصی برای زوج‌جمله‌ها و محاسبه امتیاز نهایی بین زوج‌جمله‌ها

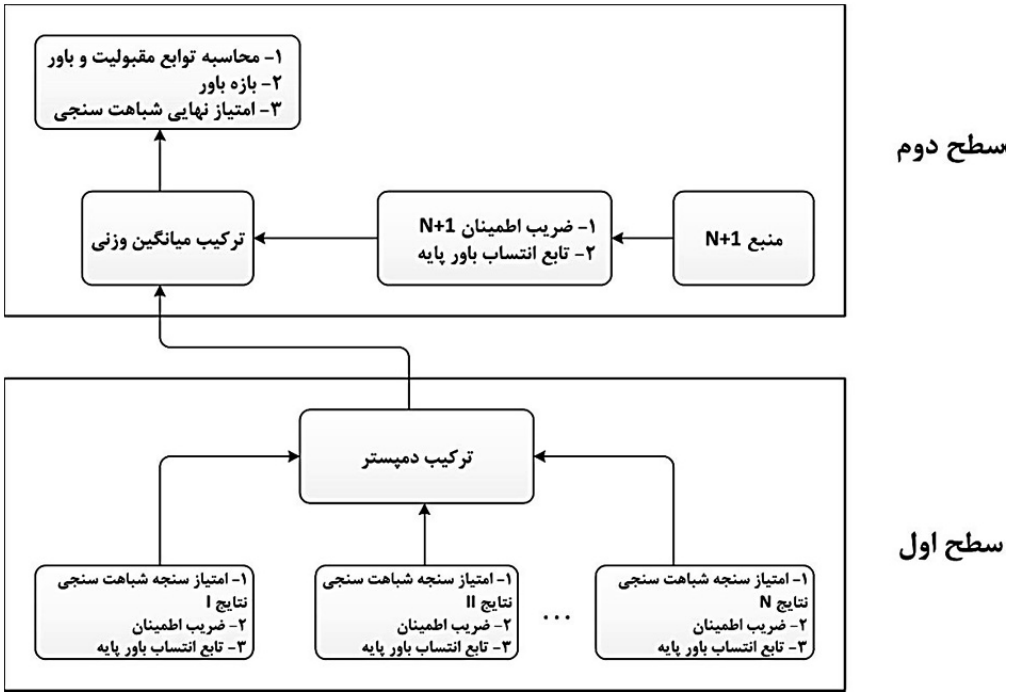
۵- تشخیص جمله‌های مشابه

باید توجه داشت که هستهٔ اصلی و قلب این سامانه «واحد محاسبهٔ شباهت» است که براساس آن پس از انجام پردازش‌های نخستین بر محتوای سند، با استفاده از سنجه‌های شباهت‌سنجی، میزان تشابه آنها اندازه‌گیری می‌شود. پژوهش‌ها نشان می‌دهد که هیچ‌یک از سنجه‌های شباهت‌سنجی در حالت کلی برتر از بقیه نیست(محمد،۲۰۰۸). این سنجه‌ها با توجه به سازوکار داخلی، به شدت وابسته به محتوای متن و ساختار طبقه‌بندی واژه‌نامه (مانند ارتفاع، ساختار و زیردرخت) هستند و کارایی و سودمندی آن با توجه به طبقه‌بندی و ساختار مفاهیم از هم متفاوت و گوناگون خواهد بود(پدرسون و همکاران،۲۰۰۴؛ ووهرینگر و فلیدل،۲۰۱۱) به همین دلیل این سنجه‌ها سنجش نادقیق و ناکاملی از زوج واژه‌ها به‌دست می‌دهند. ازهمین‌رو، در این مقاله، از چارچوب نظریهٔ گواه برای ترکیب سنجه‌های مختلف استفاده شده است. در این ایده، هر سنجه به‌عنوان یک گواه اطلاعاتی در نظر گرفته شده و سپس از نظریهٔ گواه برای پوشش معایب هر سنجه استفاده و با اعمال قانون ترکیب گواه‌ها، امتیاز نهایی برای شباهت‌سنجی دو جمله به دست می‌آید.



(شکل- ۳): معماری سامانهٔ تشخیص دستبرد ادبی

شکل (۴) جزییات انجام این فرآیند را نشان می‌دهد. در این ساختار، n سنجه در سطح نخست به‌عنوان گواه در نظر گرفته می‌شوند و سپس امتیاز حاصل از سنجش شباهت بین دو جملهٔ این n سنجه، به عنوان باور پایه برای چارچوب تشخیص نظریهٔ گواه در نظر گرفته می‌شود. این نتایج از طریق تابع هم‌جوشی دمپستر (رابطهٔ ۸) با هم ترکیب شده و به‌عنوان باور پایه به سطح دوم فرستاده می‌شود. در این مرحله برای به‌دست‌آوردن باورِ کافی نسبت به نتایج سطح قبلی، ضریب اطمینانی که از طریق کارشناسان و خبرگان منتخب تعیین شده به هر یک از گواهان (سنجه‌های مورد استفاده) تخصیص داده می‌شود. در سطح دوم نتیجهٔ حاصل از سطح نخست با باور پایهٔ حاصل از گواه سوم از طریق تابع ترکیب میانگین‌گیری (رابطهٔ ۹) با هم ترکیب و سپس این نتیجه برای محاسبهٔ تابع باور و مقبولیت (روابط ۶ و ۷) استفاده می‌شود. این توابع به عنوان بازهٔ تصمیم‌گیری برای چارچوب تشخیص در حل مسئله در نظر گرفته می‌شود و در نهایت تصمیم نهایی بر روی بهترین بازه انتخاب می‌شود و شباهت بین دو جمله به دست می‌آید.



(شکل- ۴): فرآیند هم‌جوشی سنجه‌های مختلف در بخش محاسبهٔ شباهت‌سنجی

در این سامانه، جمله‌ها به دو بخش عمومی و تخصصی بر اساس واژه‌های موجود در پایگاه دانش و هستان‌نگار[1] تقسیم و سپس زوج‌واژه‌های آنها با هم مقایسه می‌شوند. برای بخش عمومی از امتیاز ارتباط لغوی بین واژه‌ها با استفاده از واژه‌نامه عمومی استفاده شده به‌صورتی‌که چنانچه دو واژه در فهرست مترادف‌های هم باشند امتیاز یک و در غیر این صورت امتیاز صفر درنظر گرفته می‌شود و برای بخش تخصصی از روش‌های مبتنی بر پایگاه‌دانش استفاده می‌شود. در بخش تخصصی واژه‌های متن به عنوان مفهوم[2] فرض می‌شوند. این مفاهیم به‌طور معمول در هستان‌نگارها و شبکه‌های معنایی[3]، که به‌صورت ساختار درختی است، نمایش داده می‌شوند در ادامه چند روش تشابه معنایی مبتنی بر پایگاه دانش مورد استفاده در این مقاله تشریح می‌شود.

۵-۱- روش وو و پالمر[4]

در این روش تشابه معنایی با توجه به عمق (ارتفاع) دو مفهوم مورد مقایسه در طبقه‌بندی ساختار شبکه‌معنایی و همچنین مقدار عمق مفهوم مشترک بین این دو مفهوم در کمترین فاصله به‌دست می‌آید:

$$Sim\ (C_1,C_2)=\frac{\cdot\ \cdot depth\ (LCS(C_1,C_2))}{depth\,(C_1)+depth\,(C_2)} \qquad (۱۰)$$

[1] Ontology
[2] Concept
[3] Semantic networks
[4] Wu and Palmer

LCS [1] مفهوم مشترک بین دو مفهوم C_1, C_2 در ساختار درختی مفاهیم در کمترین فاصله درخت است و تابع $depth$ عمق مفهوم (ارتفاع در درخت) این مفاهیم را به‌دست می‌آورد(وو و همکاران،۱۹۹۴).

۵-۲- روش رزنیک[2]

این سنجه مبتنی بر محتوای اطلاعاتی است که مفهوم مشترک دو واژه در کمترین فاصله را در بردارد و به‌صورت رابطهٔ (۱۱) تعریف می‌شود:

$$rel(w_1, w_2) = \max[rel(C_1, C_2)] \quad C_1 \in s(w_1), C_2 \in s(w_2) \qquad (11)$$

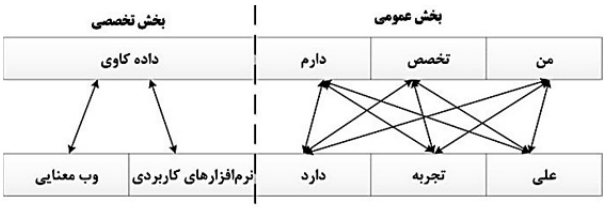
که در آن $s(w_i)$ مجموعه مفاهیمی متناظر با معانی مختلف واژهٔ w_i است. به عبارتی دیگر، C_1, C_2 مرتبط‌ترین مفاهیم در سلسله‌مراتب شبکهٔ واژگان هستند که جزو مجموعه معانی w_1, w_2 هستند. این سنجه را می‌توان به‌صورت دیگری بر مبنای محتوای اطلاعاتی تعداد فرزندان هر مفهوم در هستان‌نگار نیز تعریف کرد، که دید عینی و واقعی‌تری به دست می‌دهد. نحوهٔ محاسبه این سنجه عبارت از:

$$IC_{onto}(C) = 1 - \frac{\log(num\_desc(C)+1)}{\log(\max_{onto})} \qquad (12)$$

که $num\_desc(c)$ تعداد فرزندان هر مفهوم و $\max_{onto}$ تعداد مفهوم موجود در هستان‌نگار بوده که برای هنجارسازی مقدار تشابه به کار گرفته می‌شود(رزنیک،۱۹۹۵).

۶- پیاده‌سازی عملیاتی سامانه

سامانهٔ پیشنهادی در محیط نرم‌افزاری Visual Studio .Net و با پایگاه داده Microsoft SQL پیاده‌سازی شده و معماری طراحی سامانه به‌صورت شیءگرا[3] انجام گرفته که توانایی سنجش هر دو جمله را براساس سنجه‌های متفاوت و همچنین پایگاه دانش‌های مختلف داشته باشد. برای استفاده از نظریهٔ گواه ابتدا چارچوب تشخیص را به‌صورت فضای حالت شباهت بین دو جمله تعریف می‌کنیم و از آنجا که سنجه‌های مبتنی بر پایگاه دانش از منظرهای مختلف و با کیفیت‌های متفاوتی شباهت بین دو جمله را می‌سنجند، باعث نوعی عدم شناخت کامل در مسئله می‌شود؛ این موضوع استفاده از نظریهٔ گواه را توجیه‌پذیر می‌کند. برای کسب نتیجهٔ بهتر، هر یک از سنجه‌های مورد استفاده را به‌عنوان یک منبع (گواه) در نظر می‌گیریم و نتایج آنها را با استفاده از نظریهٔ گواه ترکیب می‌کنیم. همچنین در این مرحله با استفاده از هستان‌نگار، واژه‌های تخصصی را از واژه‌های عمومی جدا و هر جمله را به دو شبه‌جمله تقسیم می‌کنیم. برای نمونه دو جملهٔ «من در داده‌کاوی تخصص دارم» و «علی در نرم‌افزارهای کاربردی و وب‌معنایی تجربه دارد» را در نظر بگیرید؛ این جمله‌ها با استفاده از هستان‌نگار و واژه‌نامه، مطابق با شکل (۵) به دو بخش عمومی و تخصصی تقسیم و زوج‌واژه‌های آنها با هم مقایسه می‌شوند. در این مقاله از هستان‌نگار تخصصی فناوری اطلاعات (حورعلی،۱۳۹۰) استفاده شده است. یادآور می‌شود چنانچه ارتباط بین زوج‌واژه‌ای در هستان‌نگار تعریف نشده باشد؛ ابتدا از واژه‌نامهٔ عمومی برای مقایسه آنها و سپس پیشنهادی به کاربر سامانه برای گسترش هستان‌نگار داده خواهد شد.



(شکل- ۵): نحوهٔ مقایسه زوج‌واژه‌ها در زوج جمله‌ها

در این مقاله از روابط ۱۰ تا ۱۲ برای محاسبهٔ شباهت بین بخش تخصصی دو جمله استفاده شده است. در جدول ۳ مقدار شباهت زوج‌جمله بالا از طریق روابط بالا آورده شده است:

(جدول- ۳): امتیاز شباهت بین دو جمله براساس سنجه‌های مختلف

نام سنجه	امتیاز شباهت بین دو جمله
وو و پالمر (رابطه۱۰)	۰/۷۱
رزنیک (رابطه ۱۱)	۰/۹۰
رزنیک مبتنی بر تعداد گره (رابطه ۱۲)	۰/۸۴

با توجه به جدول (۳)، ملاحظه می‌شود، سنجهٔ رزنیک هم‌خوانی بهتری با نظرهای واقعی دارد و به همین دلیل برای بخش تخصصی از آن استفاده شده است. پس از

[1] Least Common Subsumer
[2] Resnik
[3] Object Oriented

انتخاب سنجهٔ شباهت‌سنجی، ماتریس‌های ۱۳ و ۱۵ را برای مقایسهٔ زوج واژه‌های دو جمله تشکیل می‌دهیم. سطر Sim_{ont} نشان‌دهندهٔ بخش تخصصی شامل واژهٔ «داده‌کاوی» جملهٔ نخست و ستون‌های آن برای جملهٔ دوم است که شامل واژه‌های «نرم‌افزارهای کاربردی ، وب‌معنایی» است:

$$Sim_{ont}(S_1, S_2) = [0.9 \quad 0.7] \qquad (۱۳)$$

پس از محاسبهٔ امتیازهای شباهت زوج‌واژه‌ها برای هر سطر و ستون ماتریس از رابطهٔ روزنفلد[1] استفاده می‌کنیم:

$$Final_{Sim_{ont}}(S_1, S_2) = \qquad (۱۴)$$
$$\frac{1}{2}\left(\sum_{i=1}^{row} Max(row) + \sum_{j=1}^{col} Max(col)\right) \forall row, Col$$

برای بخش عمومی جمله نیز ماتریس زوج‌واژه برای واژه‌های باقیماندهٔ تشکیل می‌شود که برای مثال بالا به‌صورت رابطه (۱۵) است:

$$\mathrm{Sim}_{gen}(S_1,S_2)=\begin{bmatrix}0 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix} \qquad (۱۵)$$

اینک براساس رابطهٔ ۱۶ نتیجهٔ نهایی دو بخش خصوصی و عمومی با هم ترکیب می‌شوند و به‌عنوان امتیاز شباهت سنجی نهایی دو جمله درنظر گرفته می‌شود:

(۱۶)

$$Final_{Sim}\ (S_1, S_2) = \frac{N_{ont}}{N} Final_{Sim_{ont}}(S_1, S_2) + \frac{N_{gen}}{N} Final_{Sim_{gen}}(S_1, S_2)$$

که در رابطهٔ بالا N_{gen} ،N ،N_{ont} به‌ترتیب تعداد واژه‌های موجود در هستان‌نگار ، تعداد واژه‌های عمومی و تعداد کل واژه‌ها برای دو جمله است.

هم‌چنان که بیان شد، استفاده از این سنجه‌ها به‌تنهایی سنجش ناکاملی از شباهت بین دو جمله را به دست می‌دهد. به همین دلیل در ادامه از قاعدهٔ ترکیب نظریهٔ گواه برای پوشش ناکاملی سنجش استفاده می‌کنیم. در گام نخست نتایج روابط ۱۰ و ۱۱ را با استفاده از قانون هم‌جوشی دمپستر ترکیب کرده که در جدول (۴) نمایش داده شده است. درایه‌های این جدول به صورت اشتراک طبق قاعدهٔ دمپستر به دست آمده است. به‌عنوان نمونه درایه نخست این جدول ضرب احتمال‌های دو سنجه در حالت مشابه یعنی اعداد ۰/۷۱ و ۰/۹ است که نتیجه آن به‌عنوان مشابه در نظر گرفته می‌شود. درحالتی‌که دارای ناباوری در وضعیت باشیم، نتیجه به‌صورت حالتی که مشخص است در خواهد آمد؛ به‌عنوان نمونه برای درایه نخست سطر ناباوری نتیجه مشابه در نظر گرفته شده است:

[1] Rosenfelld

جدول ۴: ترکیب باور پایه دو گواه از قاعده هم‌جوشی دمپستر

وو و پالمر (رابطه ۱۰) / زرنیک (رابطه ۱۱)	مشابه=۰/۷۱	نیمه مشابه=۰	نامشابه=۰	ناباوری=۰/۲۹
مشابه=۰/۹	= ۰/۶۳۹ مشابه	۰	۰	= ۰/۲۶۱ مشابه
نیمه مشابه=۰	۰	۰	۰	۰
نامشابه=۰	۰	۰	۰	۰
ناباوری=۰/۱۰	=۰/۰۷۱ مشابه			= ۰/۰۲۹ ناباوری
جمع کل	۰/۹۷۱	۰	۰	۰/۰۲۹

سپس نتایج سنجهٔ رابطهٔ ۱۲ از طریق ترکیب میانگین وزنی ترکیب‌شده و امتیاز نهایی شباهت‌سنجی دو جمله به‌دست می‌آید. نحوهٔ محاسبه این بخش در جدول (۵) آمده است. برای تشخیص جمله‌ها حد آستانه‌ای با توجه به شرایط متون مورد بررسی و نظر کاربر در نظر گرفته شده است. در این مقاله حد آستانه برای جمله‌های مشابه ۰/۶۵ است.

۶-۱- ارزیابی سامانه

برای ارزیابی عملکرد سامانه، از شاخص‌های «دقت»[2]، «بازخوانی»[3] و «سنجهٔ F»[4] 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. اگر A نشان‌دهندهٔ مجموعهٔ جمله‌های بازیابی‌شده و R نشان‌دهندهٔ مجموعهٔ جمله‌های مشابه برای یک جملهٔ مفروض باشند، روابط دقت و بازخوانی به‌صورت زیر است(شیخ‌اسماعیلی،۲۰۰۶):

$$\mathrm{Precision}=\frac{|A\cap R|}{|A|} \qquad (۱۷)$$

$$\mathrm{Recall}=\frac{|A\cap R|}{|R|} \qquad (۱۸)$$

این دو شاخص در جهت عکس هم حرکت می‌کنند یعنی بهبود یکی به کاهش دیگری منجر می‌شود. به‌منظور رفع این مشکل از میانگین متوازن آنها به صورت زیر استفاده می‌شود (شیخ‌اسماعیلی،۲۰۰۶):

$$F-Measure=\frac{2*\mathrm{Re}call*\mathrm{Pr}ecision}{\mathrm{Re}call+\mathrm{Pr}ecision} \qquad (۱۹)$$

[2] Percision
[3] Recall
[4] F-measure

از آنجا که هدف در این مقاله بررسی شباهت در متون تخصصی فارسی است و در حال حاضر هیچ‌گونه پیکره یا مجموعه داده‌ای با این ویژگی در زبان فارسی و همین‌طور زبان‌های دیگر وجود ندارد، از مجموعه پایگاه مقاله‌های کنفرانس یادگیری الکترونیکی برای تولید پیکره استفاده شده است. این پیکره شامل ۸۱۰ زوج‌جمله است که جملهٔ نخست زوج‌جمله‌ها از مقاله و جمله دوم توسط خبرگان زبان به‌صورت تصادفی در میزان شباهت‌های مختلف تولید شده است. به‌عنوان نمونه برای جمله «گستردگی تحولات در فنّاوری اطلاعات، موجب پدید آمدن نوع جدیدی از یادگیری به نام یادگیری الکترونیکی شده است» جمله مشابه « یادگیری الکترونیکی نوع جدیدی از یادگیری است که با استفاده از ابزارهای فناوری اطلاعات ارائه می شود» ساخته شده است. درنهایت برای ارزیابی دقیق سامانه‌ای، میزان شباهت زوج‌جمله‌ها توسط تعدادی خبره متفاوت با خبرگان تولیدکننده مشخص شده و امتیاز نهایی شباهت بین دو جمله براساس میانگین امتیاز خبرگان مدنظر قرار گرفته است.

(جدول- ۵): ترکیب نتایج با استفاده از نظریهٔ گواه

گواهان	ضریب وزنی	باور پایه			
		مشابه	نیمه مشابه	نامشابه	ناباوری
نتیجه سطح اول (روابط ۱۰ و۱۱)	۰/۵	۰/۹۷۱	۰	۰	۰/۰۲۹
نتیجه سطح دوم (رابطه ۱۲)	۰/۵	۰/۸۴	۰	۰	۰/۱۶
ترکیب باورهای پایه دو گواه از طریق قاعدهٔ میانگین‌گیری	-	۰/۹۰۵۵	۰	۰	۰/۰۹۴۵
نتیجهٔ نهایی	**دو جمله مشابه هستند**				

تاکنون هیچ‌گونه پیکرهٔ تخصصی برای تشخیص شباهت‌سنجی در متون تخصصی ارائه نشده است و بیش‌تر روش‌های ارائه‌شده در ادبیات بر روی پیکره‌های انگلیسی انجام می‌پذیرد و نمی‌توان آنها را با روش پیشنهادی مقایسه کرد. ازاین‌رو، در این مقاله نتایج پیاده‌سازی روش‌های ارائه‌شده در مقاله‌های(گوپتا و همکاران،۲۰۱۴؛موهلر و همکاران،۲۰۰۹)، فاصله ویرایشی به‌عنوان روش مستقیم و روش پیشنهادی بر روی پیکرهٔ فارسی تخصصی تولیدشده، پیاده‌سازی شده و نتایج آنها بر مبنای معیارهای F، دقت، بازخوانی و زمان با هم مقایسه شده‌اند. نتایج این مقایسه در جدول ۶ ارائه شده است. نتایج بیان‌گر آن است که روش پیشنهادی به میزان قابل قبولی، بهتر از سه روش دیگر بوده و توانایی شناسایی متون مشابه را دارد.

با توجه به نتایج جدول (۶) ملاحظه می‌شود استفاده از معماری پیشنهادی و استفاده از هم‌جوشی سنجه‌ها با کمک نظریهٔ گواه، باعث بهبود دقت در عملکرد سامانه شده و همچنین بازخوانی اسناد مرتبط و مشابه را بهتر انجام می‌دهد. از طرفی در بُعد سرعت از آنجا که در روش پیشنهادی از محاسبات بیشتری استفاده می‌شود نسبت به دیگر روش چندان تفاوت خاصی دیده نمی‌شود. اجرای این الگوریتم‌ها بر روی رایانه‌ای با CPU P4 و RAM 4g انجام شده است. لازم به ذکر است با توجه به اینکه تشخیص دستبرد ادبی به‌صورت برون‌خط انجام می‌شود، در سامانه‌های تشخیص دستبرد ادبی دقت تشخیص بیش از سرعت مدنظر است.

(جدول- ۶): نتایج کارایی روش پیشنهادی و مقایسه با سه روش دیگر

روش	معیار F	دقت	بازخوانی	سرعت اجرا (دقیقه)
روش گواه	۰/۸۶	۰/۸۴	۰/۸۷	۶/۳
روش ارائه‌شده در (موهلر و همکاران،۲۰۰۹)	۰/۷۱	۰/۶۹	۰/۷۲	۵/۴
روش ارائه‌شده در (گوپتا و همکاران،۲۰۱۴)	۰/۶۶	۰/۶۹	۰/۶۴	۷/۱
روش مستقیم (فاصله ویرایشی)	۰/۶۰	۰/۵۹	۰/۶۲	۵/۱

۷- نتیجه‌گیری و پژوهش‌های آتی

در این مقاله از نظریهٔ گواه برای هم‌جوشی اطلاعات به‌منظور شباهت‌سنجی بین دو سند و کشف دستبرد ادبی، با توجه به کیفی و ناکامل‌بودن عوامل اثرگذار سنجش شباهت بین دو متن استفاده شده است. سامانهٔ طراحی‌شده با استفاده از

گواه‌های موجود، استنتاج را در دو سطح انجام می‌دهد. در مرحلهٔ ابتدایی جمله‌های موجود در سند به دو بخش عمومی و تخصصی تقسیم و سپس با استفاده از سنجه‌های متفاوت و همچنین استفاده از هستان‌نگار تخصصی امتیاز تشابه برای هر بخش را محاسبه کرده و درنهایت امتیاز تجمعی از دو بخش برای دو جمله به دست آورده می‌شود. در سطح نخست نتایج سنجه‌های شباهت‌سنجی به‌عنوان گواه (با باور پایهٔ مشخص) با قاعدهٔ ترکیب دمپستر-شفر با هم ترکیب و به‌عنوان گواهی جدید به سطح دوم منتقل می‌شوند. در سطح دوم، نتیجهٔ سطح نخست و گواه جدید از طریق قاعدهٔ میانگین‌گیری هم‌جوش‌شده و توابع باور و مقبولیت نهایی محاسبه و شباهت بین دو جمله ارزیابی می‌شود.

این سامانه برای شباهت‌سنجی در پیکرهٔ تولیدشده توسط پژوهش‌گران که شامل ۸۱۰ زوج‌جمله است به‌کار گرفته شده است. با توجه به نتایج حاصل، ملاحظه شد سامانه در بیش از ۸۶٪ موارد، تحلیل درستی از شباهت دو جمله ارائه می‌دهد و می‌توان از این سامانه برای شناسایی مصادیق دستبرد ادبی استفاده کرد.

در حال حاضر سامانهٔ پیشنهادی با توجه به تعداد کم اسناد با سرعت خوبی پاسخ‌گوست و انتظار این است که با افزایش حجم اسناد، سامانه به سمت کندی رود که می‌توان از روش‌هایی همچون اثرانگشتی و خوشه‌بندی برای کاهش حجم سندهای مورد مقایسه استفاده کرد. از طرفی پایگاه دانش به‌کارگرفته‌شده باید غنی شود تا توانایی شناسایی ارتباط واژه‌های جدید را داشته باشد. ترکیب سنجه‌های مختلف با استفاده از روش‌هایی مانند نظریهٔ فازی می‌تواند به بهبود اندازه‌گیری شباهت معنایی و شناسایی روابط معنایی بین واژه‌ها منجر شود. از طرفی می‌توان از روش‌های شبکه‌های عصبی و ترکیب آنها با الگوریتم‌های ژنتیک برای تشخیص عبارت‌های چندواژه‌ای نیز استفاده کرد تا میزان دقت مقایسه واژه‌ها را بهبود داد.

۸- مراجع

حورعلی، مریم. یادگیری هوشمند هستان نگار برای بسط پرسمان در جستجوی معنایی، رساله دکتری، دانشگاه تربیت مدرس، ۱۳۹۰، صص ۳۰-۴۵.

Addis, Andrea. Study and Development of Novel Techniques for Hierarchical Text Categorization. Department of Electrical and Electronic Engineering. University of Cagliari,2010.

Alfred, Rayner, Leow Ching Leong, Chin Kim On, and Patricia Anthony. (2014) "A Literature Review and Discussion of Malay Rule-Based Affix Elimination Algorithms." In The 8th International Conference on Knowledge Management in Organizations, pp. 285-297. Springer Netherlands,.

Bae, h.-r., Grandhi, r. V. & Canfield, r. A. (2004).Epistemic Uncertainty quantification technique-s including evidence theory for large- scale struct-ures. Computers and Structures , 23, 125-138.

Bao J, Shen J, Liu X. (2001).On illegal copying and distributing detection mechanism for digital goods. J Comput Res Develop;38(1):121–5.

Barron-Cedeno, A. Eiselt, A. Rosso, P. Monolingual Text Similarity Measures: A Comparison of Models over Wikipedia Articles Revisions, Proc. 7th Int. Conf. on Natural Language Processing, ICON-, Hyderabad, India, pp. 29-38,2009.

Basile, C., Benedetto, D., Caglioti, E., Cristadoro, G., & Esposti, M. D. (2009). A plagiarism detection procedure in three steps: Selection, matches and "squares". In Proceedings of the 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse.

Bilenko Mikhail and Mooney Raymond J. Adaptive Duplicate Detection Using Learnable String Similarity Measures. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003), Washington DC, pp.39-48, August, 2003.

Bretag, T., & Mahmud, S. (2009). Self-plagiarism or appropriate textual re-use? Journal of Academic Ethics, 7, 193–205.

Chow,K.K. Salim,N. (2010).Web based cross language plagiarism detection, in: Second International Conference on Computational Intelligence, Modelling and Simulation, pp. 199–204.

Collberg C, Kobourov S, Louie J, Slattery T. (2003).Splat: a system for self-plagiarism detection. In: Proceedings of IADIS international conference WWW/INTERNET; p. 508–14.

Dempster, A. (1967). Upper and Lower Probabilities Induced by Multivalued Mapping, Annals of Math. Stat., 38,325-329,.

Devi, S. L., Rao, P. R. K., Ram, V. S., & Akilandeswari, A. (2010). External plagiarism dete-ction – Lab report for PAN at CLEF 2010. In Notebook Papers of CLEF LABs and Workshops.

Dreher, H. (2007).Automatic conceptual analysis for plagiarism detection. Information and Beyond: The Journal of Issues in Informing Science and Information Technology, 4, 601–614.

Ekbal A, Saha S, Choudhary G. (2012). Plagiarism detection in text using vector space model. In: 12th International conference on hybrid intelligent systems (HIS);. p. 366–71.

Frantzi, K., Ananiadou, S. , Mima. H. (2000).Automatic recognition of multi- word terms. International Journal of Digital Libraries, 3, 2, 115-130.

Geravand S, Ahmadi M. (2014).An efficient and scalable plagiarism checking system using Bloom filters. ComputElectr Eng.

M.,Mohler, and Rada Mihalcea. "Text-to-text semantic similarity for automatic short answer grading." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009.

Metzler, D. , Bernstein, Y. , Croft, W. B. , Moffat, A. , and Zobel, J. (2005).Similarity measures for tracking information flow. In Proceedings of CIKM '05, pp. 517-524.

Metzler, D. Dumais, S. Meek, C. (2007). Similarity Measures for Short Segments of Text , In Proc. of ECIR-07 Springer, Vol. 4425 , pp. 16-27 ,2007.

Meuschke, N., Gipp, B., (2013). "State-of-the-art in detecting academic plagiarism", International Journal for Educational Integrity Vol. 9 No. 1, pp. 50–71.

Mihalcea, R. Corley, C. Strapparava, C. (2006).Corpus-based and Knowledge-based Measures of Text Semantic Similarity, In American Association for Artificial Intelligence , pp. 775-780.

Mohammad, Saif. (2008).Measuring Semantic Distance Using Distributional Profiles Of Concepts, University of Toronto.

Muhr,M., Zechner, M,. Kern, R., & Granitzer, M. (2009). External and intrinsic plagiarism detection using vector space models. In Proceedings of the 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse, pp.47–55.

Osman, Ahmed Hamza, Naomie Salim, and Ammar Ahmed E. Elhadi. (2013). "A tree-based conceptual matching for plagiarism detection." In Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on, pp. 571-579. IEEE.

Osman, Ahmed Hamza. Salim, Naomie. Binwahlan, Mohammed Salem. Alteeb, Rihab. Abuobieda, Albaraa. (2012).An improved plagiarism detection scheme based on semantic role labeling, Applied Soft Computing 12 , 1493–1502.

Rahimtoroghi, E.; Faili, H.; Shakery, A, (2010). "A structural rule-based stemmer for Persian," Telecommunications (IST), 5th International Symposium on , vol., no., pp.574,578, 4-6.

Ram, R. Vijay Sundar, Efstathios Stamatatos, and Sobha Lalitha Devi. (2014)."Identification of Plagiarism Using Syntactic and Semantic Filters." In Computational Linguistics and Intelligent Text Processing, pp. 495-506. Springer Berlin Heidelberg.

Resnik, P. (1995), Using information content to evaluate semantic similarity. In Proceedings of the 14th International Joint Conference on Artificial Intelligence.

Shafer, G. (1976).A Mathematical Theory of Evidence, New Jersey, Princton University Press.

Sheykh Esmaili, k., Neshati, m., Abolhassani, a. (2006) improving ir performance using intelligent query expansion. 11 international csi computer conferences (csicc'2006).

Gipp, B., & Beel, J. (2010). Citation based plagiarism detection: A new approach to identify plagiarized work language independently. In Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (pp. 273–274).

Givi, H.A, Anvari, H. (2006). "Persian Language", 27th ed., Fatemi, Tehran.

Guan, J. W., & Bell, D. A. (1991). Evidence Theory and its applications. Amsterdam: Elsevier Science Publisher B.V.

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

Hariharan, S., Kamal, S., Faisal, A. V. M., Azharudheen, S. M., & Raman, B. (2010). Detecting plagiarism in text documents. In Proceedings of the International Conference on Recent Trends in Business Administration and Information Processing: Vol. 70. Communications in Computer and Information Science (pp. 497–500). Trivandrum, Kerala, India: Springer.

Heather, J. (2010).Turnitoff: Identifying and fixing a hole in current plagiarism detection software. Assessment & Evaluation in Higher Education, 35(6), 647–660.

Jiang, J. , and Conrath, D. (1997), Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the International Conference on Research in Computational Linguistics, pp 19–33.

Joy A.M.L.M., (1999).Plagiarism in programming assignments, IEEE Transactions on Education 42 (1) 129–133.

Khaleghi, Bahador, Alaa Khamis, Fakhreddine O. Karray, and Saiedeh N. Razavi. "Multisensor data fusion: A review of the state-of-the-art." Information Fusion 14, no. 1 (2013): 28-44.

Khatibi, V. Montazer, G.A. (2010). A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment ,Expert Systems with App-lications, 37(12): p. 8536-8542.

Kok Kent, C. , Salim, N. , Features Based Text Similarity Detection , Journal Of Computing, Volume 2, Issue 1, pp. 53-57 ,2010.

Leacock, C. , and Chodorow, M. (1998).Combining local context and WordNet sense similarity for word sense identification. In WordNet, An Electronic Lexical Database. The MIT Press.

Lesk, M. (1986).Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the SIGDOC Conference, pp. 24–26.

Li,Yuhua. McLean,David. Bandar,Zuhair A. O'Shea,James D. and Keeley Crockett .Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 8, AUGUST 2006.

حمید آهنگربهان دوره کارشناسی خـود را در سـال ۱۳۸۱ در رشـتهٔ مهندسی کامپیوتر در دانشگاه شاهد تهران گذرانده و مدرک کارشناسی ارشـد خـود را در سـال ۱۳۸۶ در رشتهٔ مهندسـی صـنایع از دانشـگاه شریف اخذ کرده است. ایشان در حال حاضر دانشـجوی دورهٔ دکترای مهندسی صنایع در دانشـگاه تربیـت مـدرس اسـت. زمینه‌های علمی مورد علاقه وی تحلیـل داده و سـامانه‌هـای اطلاعاتی با روش‌های نرم رایانش است.

نشانی رایانامهٔ ایشان عبارت است از :

Ahangarbahan@gmail.com

غلامعلی منتظر در سال ۱۳۴۸ در کازرون (فارس) به دنیا آمد. او در سال ۱۳۷۰ مدرک کارشناسی خود را در رشتهٔ مهندسی بـرق از دانشگاه صنعتی خواجه نصرالدین طوسی و سپس در سـال ۱۳۷۳ و ۱۳۷۷ به ترتیب مدارک کارشناسی ارشد و دکتـرای خـود را در همین رشته از دانشگاه تربیت مدرس اخذ کرد. وی پـس از اتمام تحصیلات به عضویت هیأت علمـی دانشـگاه تربیـت مدرس در آمد و در حال حاضـر دانشـیار مهندسـی فنـاوری اطلاعات در این دانشگاه است. حوزه‌های تخصصی وی شامل نرم رایانش (نظریه مجموعه‌های فـازی, شـبکه‌هـای عصـبی مصـنوعی, نظریـه مجموعـه‌هـای نـادقیق) و کـاربرد آن در سامانه‌های اطلاعاتی (همچون سامانهٔ یادگیری الکترونیکی و سامانه هوشمند حمل و نقل) است. وی تاکنون بـیش از ۸۰ مقالـه در نشـریات معتبـر علمـی و بـیش از ۱۹۰ مقالـه در کنفرانس‌های معتبر ملی و بین‌المللی منتشر کرده است. وی علاوه بر این موفق به دریافت جـوایز معتبـر علمـی از جملـه «برگزیده جشنواره بین‌المللی خوارزمی» ، «برنده کتاب سال دانشـگاهی ایـران»، «پژوهشـگر برگزیـده آیسسـکو» و «متخصص برجسته فناوری اطلاعات ایران» شده است.

نشانی رایانامهٔ ایشان عبارت است از

montazer@modares.ac.ir

Shivakumar,N. Garcia-Molina H., (1995).SCAM: a copy detection mechanism for digital documents, in: 2nd International Conference in Theory and Practice of Digital Libraries (DL 1995), Austin, TX, June 11–13.

Si, A., Leong, Hong, V., & Lau, R. W. H. (1997).CH-ECK: A document plagiarism detection system. In Proceedings of the ACM Symposium on Applied Computing, pp. 70–77.

Stein, B., Koppel, M., & Stamatatos, E. (Eds.). (2007) Plagiarism Analysis Authorship Identification, and Near Duplicate Detection: Vol. 276. CEUR Workshop Proceedings. CEUR-WS.org. in Proceedings of the SIGIR International Workshop, held in conjunction with the 30th Annual International ACM SIGIR Conference, Amsterdam, Netherlands. 2007.

Stein, Rosso, Stamatatos, Koppel, Agirre (Eds.): (2009).PAN'09, pp. 1-9,.

Weber-Wulff, D. (2011).Copy, Shake, and Paste – A Blog about Plagiarism written by a Professor for Media and Computing at the HTW. Online Source. Retrieved October 28, 2012, from: http://-copy-shake-paste.blogspot.com.

Weber-Wulff, D. Test cases for plagiarism detection software. (2010).In Proceedings of the 4th International Plagiarism Conference, Newcastle upon Tyne, UK.

Wu, Z., and Palmer, M., (1994), Verb semantics and lexical selection. In Proceedings of the Annual Meeting of the Association for Computational Linguistics.

Yerra., Ng.(2005).A SentenceBased Copy Detection Approach for Web Documents. Fuzzy Systems and Knowledge Discovery.

Yih, Wen-tau. Toutanova, Kristina. Platt, John C. Meek, Christopher. (2011). Learning Discriminative Projections for Text Similarity Measures. Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 247–256,Portland, Oregon, USA, 23–24 June 2011.

Yih, Wen-tau. Toutanova, Kristina. Platt, John C. Meek, Christopher. Learning Discriminative Projections for Text Similarity Measures. Proce-edings of the Fifteenth Conference on Compu-tational Natural Language Learning, pages 247–256,Portland, Oregon, USA, 23–24 June 2011.

Zhang ,J. Sun,Y. Wang, H. He,Y. Calculating Statistical Similarity between Sentences, Journal of Convergence Information Technology, Volume 6, Number 2. , pp. 22-34,2011.

Zhao Jun, Jin Qian-Li, XU Bo. (2005).Semantic Computation for Text Retrieval. Chinese Journal Of Computers, Vol. 28, No. 12, pp. 2068-2078. 12. (in Chinese).

Zini M, Fabbri M, Moneglia M, Panunzi A. (2006). Plagiarism detection through multilevel text comparison. In: Second international conference on automated production of cross media content for multi-channel distribution (AXMEDIS); p. 181–5.