Detecting Similarity in Paraphrased Persian Texts using Semantic and Probabilistic Methods

Nasrollah Pakniat*
PhD in Mathematics; Assistant Professor; Iranian Research Institute for Information Science and Technology (IranDoc);
Email: pakniat@irandoc.ac.ir

Azadeh Mohebi
PhD in System Design Engineering; Assistant Professor; Iranian Research Institute for Information Science and Technology (IranDoc) Email: mohebi@irandoc.ac.ir

Iranian Journal of Information Processing and Management

Received: 04, Mar. 2018 **Accepted: 15, Sep. 2018**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 34 | No. 4 | pp. 1823-1848
Summer 2019
https://doi.org/10.35050/JIPM010.2019.023

Abstract: Plagiarism detection is the process of locating instances of plagiarism within a work or a document. The main component of a plagiarism detection system is its text alignment algorithm aiming at detecting paraphrased passages of texts in a suspicious document, using a small set of candidate source documents. As text alignment algorithms are highly language-dependent, thus the numerous existing algorithms for other languages other than Pesian cannot be employed for Persian plagiarism detection puposes. There are different text alignment algorithms for Persian texts, while most of them are only able to detect exactly identical passages shared between texts. However, in many cases of plagiarism detection we are coping with the problem of finding similar passages that are already paraphrased. In this paper, we propose two new text alignment algorithms which are able to detect paraphrased texts in Persian language. The first one is a semantic algorithm that employs a dictionary to detect paraphrased sentences and the second one is a probabilistic algorithm that uses the statistical information obtained from a large corpus of Persian texts to detect similar texts. Compared to other existing semantic text alignment algorithms, the proposed algorithms use different measures to check the similarity between the text sentences. Furthermore, the probabilistic algorithm is the first probabilistic text alignment algorithm proposed for the Persian language. Moreover, while all existing text alignment algorithms check the similarity between any two sentences of the text separately, the proposed algorithms consider the similarity neighboring sentences in the text as well. The implementation results indicate that while the quality of both algorithms in detecting paraphrased texts is high

* Corresponding Author

enough and almost the same as each other, the proposed probabilistic method is more efficient than the proposed semantic algorithm in terms of computation time.

Keywords: Plagiarism, Semantic Text Alignment, Probabilistic Text Alignment, Paraphrased Texts

همانندجویی در متون فارسی بازنویسی‌شده با استفاده از روش‌های معنایی و احتمالاتی

نصراله پاک‌نیت
دکتری ریاضی؛ استادیار؛
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛
پدیدآور رابط pakniat@irandoc.ac.ir

آزاده محبی
دکتری؛ مهندسی طراحی سیستم‌ها؛ استادیار؛
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛
mohebi@irandoc.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۶/۱۲/۱۳ | پذیرش: ۱۳۹۷/۰۶/۲۴ | مقاله برای اصلاح به مدت ۴۹ روز نزد پدیدآوران بوده است.

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۴ | شمارهٔ ٤ | ص‌ص ۱۸۲۳-۱۸٤۸
تابستان ۱۳۹۸
https://doi.org/10.35050/JIPM010.2019.023

چکیده: همانندجویی ابزاری است که از آن برای تشخیص سرقت علمی/ادبی استفاده می‌شود. هدف در یک روش همانندجویی، تشخیص تمام قسمت‌های همانندِ موجود در یک متن مشکوک با توجه به تعدادی متن منبع احتمالی است. روش‌های زیادی برای همانندجویی ارائه شده، اما از یک طرف، استفاده از روش‌های همانندجوی موجود برای سایر زبان‌ها به‌منظور همانندجویی در زبان فارسی مناسب نیست و از طرف دیگر، اغلب روش‌های ارائه‌شده برای همانندجویی در زبان فارسی قادر به تشخیص متون بازنویسی‌شده نیستند. با توجه به این مهم، در این مقاله دو روش همانندجویی جدید با هدف تشخیص متون فارسی بازنویسی‌شده ارائه خواهد شد. روش‌شناسی پژوهش بر اساس مطالعهٔ منابع و مستندات معتبر علمی در این خصوص و روش کتابخانه‌ای است. روش اولِ پیشنهادی روشی معنایی است و از لغت‌نامه جهت بررسی همانندی جملات متون استفاده می‌کند. روش دوم پیشنهادی روشی احتمالاتی است و از اطلاعات آماری به‌دست‌آمده از پیکره‌ای عظیم از متون برای همانندجویی استفاده می‌کند. روش معنایی پیشنهادی در مقایسه با روش‌های معنایی موجود از معیارهای جدیدتری برای بررسی همانندی متون استفاده کرده و روش احتمالاتی پیشنهادی اولین روش همانندجویی احتمالاتی ارائه‌شده برای زبان فارسی است. علاوه بر این، در حالی که در سایر روش‌های موجود، همانندی هر دو جمله از متون مورد نظر به‌صورت مستقل بررسی می‌شود، در روش‌های

پیشنهادی همانندی جملات همسایه نیز در بررسی همانندی دو جمله در نظر گرفته شده است. نتایج پیاده‌سازی و آزمایشات صورت‌گرفته بر روی روش‌های پیشنهادی نشان می‌دهد که در حالی که هر دو روش از کیفیت مناسب و تقریباً یکسانی برخوردار هستند، روش همانندجوی احتمالاتیِ پیشنهادی بسیار کاراتر بوده و زمان مورد نیاز برای همانندجویی با استفاده از آن به‌طور متوسط برابر با ۳/۸ درصد زمان مورد نیاز توسط الگوریتم همانندجوی معنایی پیشنهادی است.

کلیدواژه‌ها: تشخیص سرقت علمی، همانندجویی معنایی، همانندجویی احتمالاتی، متون بازنویسی‌شده

۱. مقدمه

سرقت علمی عبارت است از استفادهٔ غیرمجاز یا تقلید نزدیک از ایده‌ها و نگارش شخصی دیگر و ارائهٔ آن به‌عنوان کار خود (Hannabuss 2001). همانندجویی ابزاری است که از آن برای تشخیص سرقت علمی استفاده می‌شود. هر روش همانندجویی از سه فرایند تشکیل شده است: ۱) بازیابی منبع، ۲) تطبیق متن، و ۳) پردازش نهایی. هدف در فرایند بازیابی منبع، با فرض در اختیار داشتن یک متن علمی مشکوک، بازیابی تمام منابعی است که از متن آن‌ها در نگارش متن مشکوک استفاده شده است؛ به‌طوری که هزینه‌های بازیابی کمینه شود. در فرایند تطبیق متن با ورودی یک متن مشکوک و یک متن منبع احتمالی، این متون به واحدهایی (مانند پاراگراف، جمله یا دنباله‌های چندتایی از لغات و کاراکترها) تقسیم‌بندی شده و همانندی هر دو واحد مورد بررسی واقع می‌شود. هدف این فرایند مشخص‌نمودن همهٔ قسمت‌های متن مشکوک است که با استفاده از متن منبع بازنویسی‌شده‌اند. هدف در فرایند پردازش نهایی، با ورودی متن مشکوک انجام بررسی‌های بیشتر برای ویرایش‌های نهایی در قسمت‌هایی است که با متن منبع احتمالی مشابه تشخیص داده شده است (Potthast et al. 2013).

در عمل، اولین فرایند همانندجویی یعنی بازیابی منبع با استفاده از روش‌های بازیابی اطلاعات انجام می‌شود. برای این منظور، از روش‌های معمول بازیابی اطلاعات استفاده می‌شود که می‌توانند بر اساس عملگرهای بولین، روش‌های مبتنی بر مدل‌های فضای برداری، روش‌های مبتنی بر هش و اثر انگشت، و روش‌های مبتنی بر الگوریتم‌های Word2Vec و doc2Vec باشند (Alzahrani, Salim and Abraham 2012; Mikolov et al. 2013; Le and Mikolov 2014). با توجه به این مهم، در سالیان اخیر این فرایند از مسئلهٔ همانندجویی

جدا شده و تنها فرایندهای تطبیق متن و پردازش نهایی به‌عنوان فرایندهای همانندجویی در نظر گرفته می‌شوند. در این پژوهش تمرکز روی ارائهٔ روشی برای همانندجویی متون علمی نگارش‌شده به زبان فارسی است. روش‌ها و نرم‌افزارهای زیادی برای همانندجویی متون نوشته شده و در زبان‌های مختلف ارائه شده است که بسیاری از آن‌ها وابسته به قواعد و ویژگی‌های خاص هر زبان هستند و نمی‌توان از آن‌ها به‌راحتی برای زبان فارسی استفاده کرد. از طرف دیگر، اغلب تحقیقات انجام‌شده در زمینهٔ همانندجویی در متون فارسی تنها قادر به تشخیص قسمت‌های دقیقاً یکسان از متن بوده و نمی‌توانند قسمت‌های بازنویسی‌شده از متن را تشخیص دهند. بنابراین، مسئله‌ای که پژوهش حاضر به حل آن می‌پردازد، یافتن روشی است که از طریق آن بتوان همانندی متون بازنویسی‌شده را محاسبه نمود. با توجه به این مهم، در این مقاله دو روش همانندجویی جدید برای زبان فارسی با هدف تشخیص متون بازنویسی‌شده ارائه می‌شود. روش اولِ پیشنهادی، یک روش معنایی است که برای بررسی همانندی دو قسمت مختلف متن از لغت‌نامه استفاده می‌کند. در مقایسه با سایر روش‌های معناییِ ارائه‌شده برای زبان فارسی، در این روش از معیارهای جدیدتری برای بررسی همانندی استفاده می‌شود. علاوه بر این، در این روش، همانندی دو واحدِ مختلف متن به‌صورت مجزا بررسی نمی‌شود، بلکه در بررسی همانندی این دو واحد، همانندی واحدهای نزدیک به آن‌ها در متن مشکوک و متن منبع احتمالی نیز در نظر گرفته می‌شود. روش دوم پیشنهادی، اولین روش همانندجوی احتمالاتی ارائه‌شده برای زبان فارسی است که در آن با استفاده از اطلاعات آماری به‌دست‌آمده از پیکره‌ای عظیم از متون فارسی، همانندی دو واحد متن بررسی می‌شود. همانند روش معنایی پیشنهادی، روش احتمالاتی پیشنهادی نیز در بررسی همانندی دو واحد متن، همانندی واحدهای مجاور را در نظر می‌گیرد.

در ادامهٔ این مقاله، در بخش دوم پیشینهٔ مسئله بررسی می‌شود. در بخش سوم، به روش‌های همانندجوی پیشنهادی اشاره شده و در بخش چهارم، به ارزیابی روش‌های ارائه‌شده خواهیم پرداخت. در نهایت، نتیجه‌گیری در بخش پنجم بیان می‌شود.

۲. پیشینهٔ پژوهش

در این بخش، ابتدا به بررسی روش‌های ارائه‌شده برای همانندجویی در سایر زبان‌ها پرداخته و سپس، روش‌های ارائه‌شده برای همانندجویی در زبان فارسی را بررسی خواهیم کرد.

۲-۱. همانندجویی در سایر زبان‌ها

روش‌های همانندجویی تک‌زبانی را می‌توان با توجه به ابزارهای مورد استفاده در آن‌ها به دسته‌های ۱) همانندجویی مبتنی بر کاراکتر، ۲) همانندجویی مبتنی بر بردار، ۳) همانندجویی مبتنی بر گرامر، ۴) همانندجویی معنایی، ۵) همانندجویی احتمالاتی، ۶) همانندجویی ساختاری، و ۷) همانندجویی مبتنی بر سبک نگارش تقسیم‌بندی کرد (Alzahrani, Salim and Abraham 2012.). روش‌های همانندجویی مبتنی بر سبک نگارش فقط یکتایی نویسنده در یک متن را بررسی کرده و از دقت بالا در تشخیص قسمت‌های همانند برخوردار نیستند. روش‌های همانندجویی ساختاری فقط در تشخیص سرقت ایده کاربرد دارند. روش‌های مبتنی بر کاراکتر، روش‌های مبتنی بر بردار و روش‌های گرامری در تشخیص همانندی‌های حاصل از بازنویسی متون از کیفیت مناسبی برخوردار نیستند. از میان دسته‌های بیان‌شده تنها دسته‌های شماره ۴ و ۵ یعنی روش‌های معنایی و احتمالاتی به نحوی مطلوب قادر به تشخیص همانندی‌های حاصل از بازنویسی متون هستند. در ادامهٔ این بخش، به بررسی روش‌های همانندجویی معنایی و احتمالاتی ارائه‌شده برای سایر زبان‌ها می‌پردازیم. لازم به ذکر است که در کلیهٔ روش‌هایی که در ادامه بررسی می‌کنیم، ابتدا پیش‌پردازش‌های خاص زبانی اعمال می‌شود.

در روش ارائه‌شده توسط Li et al. (2006)، که یک روش همانندجویی معنایی برای همانندجویی در زبان انگلیسی است، برای بررسی همانندی، ابتدا لغات وزن‌دهی‌شده و سپس با استفاده از یک لغت‌نامه، همانندی مابین لغات به همانندی بین جملات توسعه داده می‌شود. علاوه بر این، روش مورد نظر برای دستیابی به نتایج بهتر همانندی ترتیبی حضور لغات را نیز در نظر می‌گیرد. در پژوهش Adam (2014)، به‌طور همزمان از ابزار گرامری «تشخیص‌دهندهٔ اقسام کلمه» و لغت‌نامه برای همانندجویی استفاده شده است. Yerra and Ng (2005) یک روش همانندجویی احتمالاتی برای زبان انگلیسی ارائه کرده‌اند. در این روش، با استفاده از پیکره‌ای بزرگ و با توجه به هم‌رخدادی هر دو واژه در متون، مقداری احتمالاتی بین ۰ و ۱ به‌عنوان همانندی آن‌ها اختصاص داده شده و سپس، میانگین بیشینهٔ همانندی لغات جملهٔ اول با لغات جملهٔ دوم به‌عنوان همانندی دو جمله محاسبه شده است. در این روش، برای محاسبهٔ همانندی هر دو واژه تنها حضور آن‌ها در متون مختلف در نظر گرفته می‌شود و تعداد تکرار آن‌ها در متون در نظر گرفته نمی‌شود. در روش Koberstein and Ng (2006)، با در نظر گرفتن دفعات تکرار واژگان در متون، روش ارائه‌شده برای تعیین

میزان همانندی هر دو واژه در Yerra and Ng (2005)، در راستای دستیابی به نتایج بهتر اصلاح شده است. با وجود این، همان‌طور که نویسندگان بیان کرده‌اند، بهبود به‌دست‌آمده در نتایج این روش در مقایسه با افزایش قدرت محاسباتی مورد نیاز آن قابل توجه نیست. Alzahrani and Salim (2009)، روشی مشابه با روش ارائه‌شده در Yerra and Ng (2005) برای زبان عربی ارائه کرده‌اند. در روش Gipp and Meuschke (2011)، از روش‌های همانندجویی معنایی در کنار الگوریتم‌های تطبیق رشته برای همانندجویی استفاده شده است. Alzahrani, Salim and Palade (2015)، با تلفیق روش‌های همانندجویی معنایی و احتمالاتی ارائه‌شده در Li et al. (2006) و Yerra and Ng (2005) و همچنین استفاده از تشخیص‌دهنده‌های اقسام کلمه نتایج بهتری برای همانندجویی در زبان انگلیسی به‌دست آورده‌اند.

۲-۲. همانندجویی در زبان فارسی

اغلب روش‌های همانندجویی ارائه‌شده تا به امروز برای زبان فارسی تنها جهت تشخیص کپی دقیق بوده‌اند که از مهم‌ترین آن‌ها می‌توان به Mahmoodi and Varnamkhasti (2014); Mahdavi, Siadati and Yaghmaee (2014) Rafieian (2016); و Minaei and Niknam (2016) اشاره کرد. در روشی که Gharavi et al. (2016) ارائه کرده‌اند، از روش‌های یادگیری عمیق برای همانندجویی استفاده شده است. در این روش با استفاده از روش‌های شبکهٔ عصبی عمیق، هر لغت با یک بردار نمایش داده شده که این بردار حاوی اطلاعات معنایی و گرامری در مورد آن لغت است که در نهایت، از این بردارها برای بررسی همانندی دو متن استفاده شده است. لازم به ذکر است که این روش برای مسابقات همانندجویی در زبان فارسی طراحی شده و جزئیات چندان دقیقی دربارهٔ آن بیان نشده است. Mashhadirajab and Shamsfard (2016) یک روش همانندجوی جدید ارائه کرده‌اند که می‌توان آن را به‌عنوان یک روش ترکیبی از روش‌های برداری و معنایی در نظر گرفت. این روش جهت ارائه به اولین دورهٔ مسابقات همانندجویی طراحی شده است. با توجه به این موضوع، از آنجا که در مجموعهٔ متون مسابقه در هر متن تنها یک نوع سرقت علمی از ۳ نوع سرقت علمی (کپی دقیق، شبیه‌سازی کامپیوتری و شبیه‌سازی انسانی) انجام شده، در این روش، ابتدا با استفاده از روش‌های دسته‌بندی، نوع سرقت علمی مشخص شده و سپس، از شبکهٔ واژگانی زبان فارسی برای بررسی همانندی دو متن استفاده می‌شود. با توجه به این‌که در دنیای واقعی سرقت علمی به تفکیکِ در نظر گرفته‌شده در Mashhadirajab and Shamsfard

(2016) انجام نمی‌شود، به نظر می‌رسد پیاده‌سازی این روش در دنیای واقعی به نتایج ضعیف‌تری منجر گردد.

با توجه به ادبیات موضوع و پژوهش‌های پیشین، روش‌های متنوعی برای همانندجویی متون بازنویسی‌شده در زبان‌های دیگر ارائه شده است، لیکن بسیاری از آن‌ها به‌سادگی قابل پیاده‌سازی برای زبان فارسی نبوده و به قواعد زبانی وابسته هستند. از طرفی، از بین روش‌های ارائه‌شده برای زبان فارسی، روشی که Gharavi et al. (2016) ارائه کرده‌اند، نتایج نسبتاً مشابهی با سایر روش‌ها دارد و بر اساس نتایج آزمایش‌های انجام‌شده در پژوهش ایشان، برخی شاخص‌ها مانند دقت و سرعت از سایر روش‌ها بهتر است. در روش آن‌ها بُرداری برای هر جمله بر اساس میانگین کلمات آن بیان می‌شود و کلمات دو جمله به‌صورت جداگانه با هم مقایسه نمی‌شوند. بنابراین، ممکن است با میانگین‌گیری، اثر کلمات مشابه از بین برود. در پژوهش حاضر، روشی پیشنهاد شده است که می‌تواند در سطح کلمه، همانندی بین جملات را بررسی کند.

۳. روش پژوهش

در این پژوهش با مطالعهٔ منابع علمی معتبر دو روش پیشنهاد می‌شود که از طریق آن‌ها بتوان همانندی دو متن بازنویسی‌شده را محاسبه نمود. برای جمع‌آوری اطلاعات پژوهش، از روش کتابخانه‌ای استفاده شده و نتایج به‌کارگیری دو روش پیشنهادی روی مجموعه‌ای از داده‌ها ارزیابی و تحلیل شده است. برای تحلیل نتایج و مقایسهٔ آن‌ها از شاخص‌های مختلفی نظیر دقت، فراخوانی، دانه‌دانه بودن و *plagdet* استفاده شده است.

پیاده‌سازی روش‌های ارائه‌شده با استفاده از زبان برنامه‌نویسی c# انجام شده و در ادامه، از کامپیوتری با پردازندهٔ مرکزی "3.7 GHz Intel i3"، حافظهٔ داخلی "4 GB" و سیستم‌عامل ۶۴ بیتی ویندوز ۷ آزمایشاتی جهت بررسی کیفیت الگوریتم‌های پیشنهادی استفاده شده است. آزمایش‌های صورت‌گرفته بر روی مجموعه داده آزمایش منتشر شده با این هدف و برای کنفرانس تشخیص سرقت علمی در زبان فارسی انجام شده است. در ادامه، برخی ملاحظات صورت‌گرفته در پیاده‌سازی الگوریتم‌های پیشنهادی بیان می‌شود:

- ◇ در پیاده‌سازی‌های انجام‌شده، برای ریشه‌یابی لغوی از «ابزار پارسر زبان فارسی» (استیری و همکاران ۱۳۹۱) استفاده شده است.
- ◇ لیست ایست‌واژه‌های زبان فارسی مورد استفاده در پیاده‌سازی‌های صورت‌گرفته با

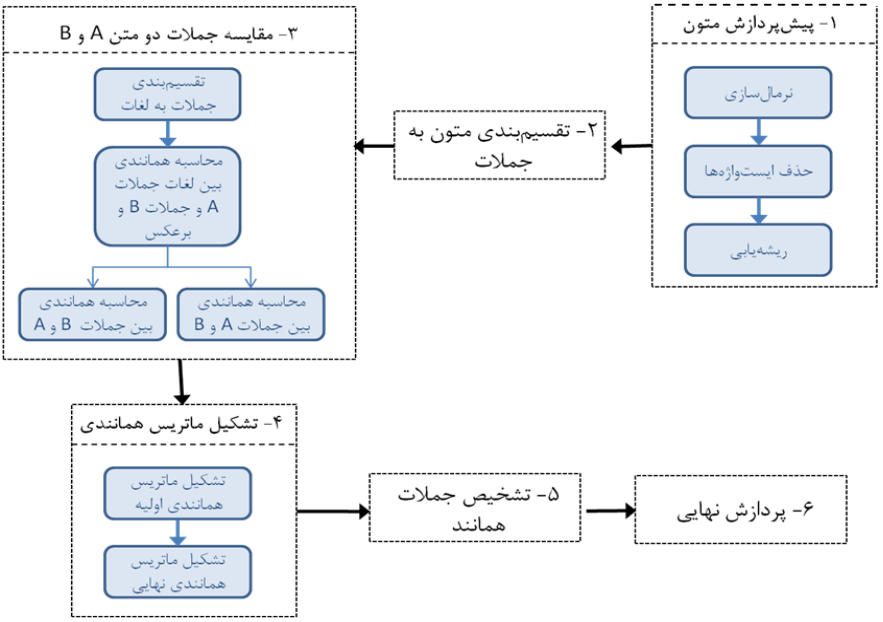
توجه به پژوهش Taghva, Beckley and Sadeh (2003) به‌دست آمده است.

- لغت‌نامهٔ مورد استفاده در الگوریتم معنایی پیشنهادی جهت بررسی نزدیکی معنایی دو واژه «فرهنگ جامع واژگان مترادف و متضاد زبان فارسی» تألیف Khodaparasti (1997) است.
- برای محاسبهٔ فراوانی واژگان و فراوانی هم‌رخدادی واژگان به‌منظور استفاده در الگوریتم همانندجوی احتمالاتی پیشنهادی، با توجه به قدرت محاسباتی در دسترس، از قسمتی از پیکرهٔ «همشهری» AleAhmad et al. (2009) به‌عنوان پیکرهٔ مرجع استفاده شده (حدود ۲۰ درصد از داده‌های این پیکره که به‌صورت تصادفی انتخاب شده‌اند) و جدول هم‌رخدادی لغات در هر دنباله ۳۱ کلمه‌ای پیاپی از متون این مجموعه محاسبه شده است.

لازم به ذکر است که متأسفانه با توجه به عدم دسترسی به پیاده‌سازی سایر روش‌های ارائه‌شده برای همانندجویی در زبان فارسی، عدم دسترسی به مجموعهٔ دادهٔ مورد استفاده برای آزمایش این روش‌ها و همچنین، نبودِ نتایج ارزیابی این روش‌ها روی مجموعهٔ دادهٔ مورد بررسی در این مقاله، امکان مقایسهٔ روش‌های پیشنهادی با این روش‌ها ممکن نیست.

۳-۱. روش‌های پیشنهادی برای همانندجویی

در این بخش، دو روش همانندجویی جدید برای زبان فارسی ارائه می‌شود. روش اولِ پیشنهادی یک روش معنایی است که از ابزار لغت‌نامه برای بررسی همانندی واحدهای مختلف متن استفاده می‌کند. روش دومِ پیشنهادی یک روش احتمالاتی است که در آن همانندی هر دو واحد از متن با توجه به اطلاعات آماری به‌دست‌آمده از پیکره‌ای عظیم از متون فارسی محاسبه می‌شود. فرایند اصلی هر دو روش در شکل ۱، نمایش داده شده است. در ادامه، هر یک از بخش‌های این فرایند برای هر دو روش پیشنهادی به تفکیک تشریح می‌گردد.



شکل ۱. فرایند کلی در هر دو روش همانندجویی پیشنهادی

۳-۱-۱. روش همانندجویی معنایی پیشنهادی

در روش همانندجوی معنایی پیشنهادی، ابتدا پیش‌پردازش‌های ریشه‌یابی لغوی، نرمال‌سازی متن و حذف ایست‌واژه‌ها بر روی متن مشکوک و متن منبع احتمالی انجام می‌شود. در ادامه، پس از تقسیم متون ورودی به واحدهای مناسب، همانندی هر دو واحد با استفاده از ابزار لغت‌نامه تعیین می‌شود. در مقایسه با سایر روش‌های معنایی ارائه‌شده برای زبان فارسی، در این روش از معیارهای جدیدتری برای بررسی همانندی استفاده می‌شود. علاوه بر این، در این روش، همانندی دو واحد مختلف متن به‌صورت مجزا بررسی نمی‌شود، بلکه در بررسی همانندی این دو واحد، همانندی واحدهای نزدیک به آن‌ها در متن مشکوک و متن منبع احتمالی نیز در نظر گرفته می‌شود.

فرض کنید A متن مشکوک و B متن منبع احتمالی باشد. در روش معنایی پیشنهادی برای مقایسهٔ A و B به‌صورت زیر عمل می‌شود:

پیش‌پردازش متون: پیش‌پردازش‌های زیر بر روی متن مشکوک A و متن منبع احتمالی B انجام می‌شود:

نرمال‌سازی: در نرمال‌سازی، تنها حروف فارسی و علائم نقطه‌گذاری نگهداری شده و بقیهٔ کاراکترها حذف می‌شوند.

حذف ایست‌واژه‌ها: با توجه به فراوانی استفاده از ایست‌واژه‌ها، این پیش‌پردازش از تأثیر منفی ایست‌واژه‌ها در همانندجویی جلوگیری می‌کند.

ریشه‌یابی: ریشه‌یابی لغوی بر روی هر دو متن اعمال می‌شود تا لغات متون به ریشه‌های آن‌ها در زبان فارسی تبدیل شوند.

تقسیم‌بندی متون به جملات: متون A و B به جملات تقسیم‌بندی می‌شوند. در این‌صورت AS_1، AS_2، ...، $AS_{|A|}$ جملات موجود در متن A و BS_1، BS_2، ...، $BS_{|B|}$ جملات موجود در متن B به‌ترتیبِ حضور در این متون و $|A|$ و $|B|$ تعداد جملات متون مورد نظر هستند.

مقایسهٔ جملات دو متن: هر جملهٔ AS_i از متن A با هر جملهٔ BS_j از متن B به‌صورت زیر مقایسه می‌شود:

تقسیم‌بندی جملات به لغات: جملات AS_i و BS_j به مجموعهٔ لغات تقسیم‌بندی می‌شوند. فرض کنیم $\{AS_i^1, AS_i^2, ..., AS_i^n\}$ مجموعهٔ لغات جملهٔ AS_i و $\{BS_j^1, BS_j^2, ..., BS_j^m\}$ مجموعهٔ لغات جملهٔ BS_j باشند.

محاسبهٔ همانندی بین لغات جمله‌ای از متن A با جمله‌ای از متن B و برعکس: همانندی هر لغت AS_i^k از جملهٔ AS_i و هر لغت BS_j^l از جملهٔ BS_j که با $Sim_{AS_i^k \leftrightarrow BS_j^l}$ نمایش داده می‌شود، به‌صورت زیر محاسبه می‌شود:

- ◇ در صورتی که دو واژه یکسان باشند، $Sim_{AS_i^k \leftrightarrow BS_j^l}$ برابر با ۱ در نظر گرفته می‌شود.
- ◇ در غیر این صورت، مقدار $Sim_{AS_i^k \leftrightarrow BS_j^l}$ با استفاده از لغت‌نامهٔ مورد استفاده در الگوریتم محاسبه می‌شود. در آزمایشات صورت‌گرفته از لغت‌نامه‌ای عمومی، که تنها شامل لغات مترادف و متضاد است، استفاده شده و در نتیجه، در صورت وجود یکی از این کلمات مابین مترادف‌ها و متضادهای دیگر مقدار $Sim_{AS_i^k \leftrightarrow BS_j^l}$ برابر با γ قرار خواهد گرفت (مقدار γ توسط آزمایش‌ها محاسبه می‌شود).

در روش‌های معنایی، برای محاسبهٔ همانندی یک لغت به یک جمله، همانندی نزدیک‌ترین لغت از جمله به لغت مورد نظر در نظر گرفته می‌شود. بنابراین، در روش پیشنهادی،

همانندی هر لغت AS_i^k از جملهٔ AS_i با جملهٔ BS_j که با $Sim_{AS_i^k \leftrightarrow BS_j}$ نمایش داده می‌شود، برابر با بیشینهٔ $Sim_{AS_i^k \leftrightarrow BS_j^l}$ برای $l = 1,...,m$ در نظر گرفته می‌شود.
به همین ترتیب، همانندی هر واژهٔ BS_j^l از جملهٔ BS_j با جملهٔ AS_i محاسبه شده و با $Sim_{BS_j^l \leftrightarrow AS_i}$ نمایش داده می‌شود.

محاسبهٔ همانندی هر جمله از متن A با هر جمله از متن B بر اساس همانندی لغات آن‌ها: میزان همانندی جملهٔ AS_i نسبت به جملهٔ BS_j با استفاده از فرمول زیر محاسبه می‌شود:

(۱)
$$Sim_{AS_i \to BS_j} = \frac{\sum_{k=1}^{n} Sim_{AS_i^k \to BS_j}}{n}$$

محاسبهٔ همانندی هر جمله از متن B با هر جمله از متن A بر اساس همانندی لغات آن‌ها: میزان همانندی جملهٔ BS_j نسبت به جملهٔ AS_i با استفاده از فرمول زیر محاسبه می‌شود:

(۲)
$$Sim_{BS_j \to AS_i} = \frac{\sum_{l=1}^{m} Sim_{BS_j^l \to AS_i}}{m}$$

تشکیل ماتریس همانندی: ابتدا ماتریس همانندی اولیه تشکیل می‌شود. سپس، با در نظر گرفتن اثر جملات همسایه، این ماتریس بهبود می‌یابد و ماتریس همانندی نهایی تشکیل می‌شود.

تشکیل ماتریس‌های همانندی اولیه: با استفاده از مقادیر محاسبه‌شده به‌عنوان میزان همانندی جملات مختلف دو متن، ماتریس‌های اولیه $Sim_{A \to B} = \left[a_{ij}\right]_{i=1...|A|}^{j=1...|B|}$ و $Sim_{B \to A} = \left[b_{ij}\right]_{i=1...|A|}^{j=1...|B|}$ ساخته می‌شوند که

(۳)
$$a_{ij} = \begin{cases} 2 \; if \quad Sim_{AS_i \to BS_j} > \alpha_1 \;\&\; \left(Sim_{AS_i \to BS_j}\right)^{\frac{1}{\log_{\delta}^{n/\delta}}} > \alpha_1' \\ 1 \; if \quad \beta_1 < Sim_{AS_i \to BS_j} < \alpha_1 \;\&\; \beta_1' < \left(Sim_{AS_i \to BS_j}\right)^{\frac{1}{\log_{\delta}^{n/\delta}}} < \alpha_1' \\ 0 \; if \quad Otherwise \end{cases}$$

و

$$b_{ij} = \begin{cases} 2 \; if \quad Sim_{BS_j \to AS_i} > \alpha_2 \;\&\; \left(Sim_{BS_j \to AS_i}\right)^{\frac{1}{\log_\delta^{m/\delta}}} > \alpha_2' \\ 1 \; if \quad \beta_2 < Sim_{BS_j \to AS_i} < \alpha_2 \;\&\; \beta_2' < \left(Sim_{BS_j \to AS_i}\right)^{\frac{1}{\log_\delta^{m/\delta}}} < \alpha_2' \\ 0 \; if \quad Otherwise \end{cases} \quad (۴)$$

که α_1، α_1'، α_2، α_2'، β_1، β_1'، β_2 و β_2' مقادیری آستانه‌ای و δ ثابت بوده و با انجام آزمایش محاسبه می‌شوند. همان‌طور که از روابط (۱) و (۲) مشخص است، طول جملات یعنی n و m در محاسبهٔ مقادیر از طریق این فرمول‌ها تأثیر زیادی دارند. این اثرگذاری زمانی باعث مشکل خواهد شد که جمله‌ها در بازنویسی، ترکیب یا تجزیه شوند. برای حل این مشکل باید به طریقی این اثرگذاری را در نظر گرفت. به‌عنوان یک راهکار، می‌توان از مقادیر آستانه‌ای متفاوت برای جملات با طول متفاوت استفاده کرد. به‌عنوان یک روش کلی‌تر، در روش ارائه‌شده در این مقاله، از یک رابطهٔ ابتکاری استفاده شده که طی آن، با به‌توان $\frac{1}{\log_\delta^{n/\delta}}$ رساندن مقادیر حاصل از فرمول‌های (۱) و (۲) این موضوع در نظر گرفته شده است. توجه کنید که از آنجا که مقادیر همانندی محاسبه‌شده از طریق فرمول‌های (۱) و (۲) بین صفر و یک هستند، با به توان $\frac{1}{\log_\delta^{n/\delta}}$ رساندن، مقدار ثانویهٔ محاسبه‌شده برای جملات طولانی افزایش خواهد یافت. برای مثال، با در نظر گرفتن δ=2، مقدار $\frac{1}{\log_\delta^{n/\delta}}$ برای n=4 برابر با ۱، برای n=8 این مقدار برابر با $\frac{1}{2}$، و برای n=16 این مقدار برابر با $\frac{1}{4}$ خواهد بود. در نتیجه، برای nهای بزرگ‌تر، استفاده از این رویکرد منجر به افزایش مقدار محاسبه‌شده خواهد شد و بدین طریق، ترکیب و تجزیهٔ جملات تا حدودی در نظر گرفته خواهد شد.

تشکیل ماتریس همانندی نهایی: با استفاده از مؤلفه‌های دو ماتریس به‌دست‌آمده در مرحلهٔ قبل، ماتریس نهایی همانندی $H = [h_{ij}]$ به‌صورت زیر ایجاد می‌شود:

$$h_{ij} = \begin{cases} 1 \quad if \;\; \left(a_{ij} \geq 1 \& b_{ij} \geq 1\right) \;\&\; \left(\exists k,l \in \{-1,0,1\} s.t.: \; (a_{i+k\; j+l} = 2 \;\&\; b_{i+k\; j+l} = 2)\right) \\ 0 \quad otherwise \end{cases} \quad (۵)$$

تشخیص جملات همانند: در نهایت، جملهٔ AS_i از متن A و جملهٔ BS_j از متن B همانند

شناخته می‌شوند اگر $h_{ij}=1$.

پردازش نهایی: پس از مشخص‌شدن جملات مشابه دو متن، فرایند زیر بر روی متن مشکوک و متن منبع احتمالی انجام می‌شود تا قسمت‌های همانند تشخیص داده‌شدهٔ مجزا به یکدیگر متصل شده و قسمت‌های با طول کم حذف شوند:

در صورتی که دو جملهٔ نزدیک به هم در متن مشکوک (در متن منبع احتمالی) همانند تشخیص داده شود و فاصلهٔ آن‌ها در متن کمتر از مقداری آستانه‌ای باشد، قسمت همانند تشخیص داده‌شده، از ابتدای جملهٔ اول تا انتهای جملهٔ دوم در نظر گرفته می‌شود. در پیاده‌سازی‌های صورت‌گرفته، از نصف مجموع طول دو جمله همانند تشخیص داده‌شده به‌عنوان این مقدار آستانه‌ای استفاده شده است.

در صورتی که قسمتی از متن همانند تشخیص داده شده و پس از انجام مرحلهٔ قبل طول آن همچنان کمتر از مقدار آستانه‌ای باشد، این قسمت از مجموعهٔ متون همانند تشخیص داده‌شده حذف می‌گردد. در پیاده‌سازی‌های صورت‌گرفته از ۴ کلمه به‌عنوان این مقدار آستانه‌ای استفاده شده است.

توجه: در انجام سرقت علمی، به‌طور معمول، بخش‌هایی شامل چندین جملهٔ پیاپی از یک متن منبع در متنی جدید بازنویسی می‌شود. با توجه به این مهم، در صورت تشخیص همانندی مابین یک جمله از یک متن مشکوک و یک جمله از یک متن منبع، در بررسی همانندی جملات مجاور آن‌ها باید حساسیت بیشتری قائل شد. همان‌طور که نتایج پیاده‌سازی روش‌های پیشنهادی (در بخش ۴) نشان می‌دهد، این اقدام (که از طریق مرحلهٔ تشکیل ماتریس همانندی نهایی در الگوریتم فوق و الگوریتم همانندجویی احتمالاتی ارائه‌شده در بخش بعد پیاده‌سازی شده) منجر به دستیابی به نتایجی بهتر شده است.

۳-۱-۲. روش همانندجویی احتمالاتی پیشنهادی

در این قسمت یک روش احتمالاتی برای همانندجویی در متون فارسی بازنویسی‌شده ارائه می‌کنیم. همانند الگوریتم ارائه‌شده در بخش قبل، در این روش نیز ابتدا پیش‌پردازش‌های ریشه‌یابی لغوی، نرمال‌سازی متن و حذف ایست‌واژه‌ها بر روی هر دو متن ورودی انجام می‌شود. در ادامه، پس از تقسیم متون ورودی به واحدهای مناسب، همانندی هر دو واحد با استفاده از فراوانی‌های لغات و فراوانی‌های هم‌رخدادی آن‌ها

با توجه به پیکره‌ای عظیم از متون فارسی محاسبه می‌شود. همانند روش قبل، در روش احتمالاتی پیشنهادی نیز همانندی دو واحد مختلف متن به‌صورت مجزا بررسی نمی‌شود، بلکه در بررسی همانندی این دو واحد، همانندی واحدهای نزدیک به آن‌ها در متن مشکوک و متن منبع احتمالی نیز در نظر گرفته می‌شود. فرض کنید A متن مشکوک و B متن منبع احتمالی باشد. در روش احتمالاتی پیشنهادی برای مقایسهٔ یک متن مشکوک با یک متن منبع احتمالی به‌صورت زیر عمل می‌شود:

پیش‌پردازش متون: پیش‌پردازش‌های زیر بر روی متن مشکوک A و متن منبع احتمالی B انجام می‌شود:

نرمال‌سازی: در نرمال‌سازی تنها حروف فارسی و علائم نقطه‌گذاری نگهداری شده و بقیهٔ کاراکترها حذف می‌شوند.

حذف ایست‌واژه‌ها: با توجه به فراوانی استفاده از ایست‌واژه‌ها، این پیش‌پردازش از تأثیر منفی ایست‌واژه‌ها در همانندجویی جلوگیری می‌کند.

ریشه‌یابی: ریشه‌یابی لغوی بر روی هر دو متن اعمال می‌شود تا لغات متون به ریشه‌های آن‌ها در زبان فارسی تبدیل شوند.

تقسیم‌بندی متون به جملات: متون A و B به جملات تقسیم‌بندی می‌شوند. در این‌صورت AS_1، AS_2، ...، $AS_{|A|}$ جملات موجود در متن A و BS_1، BS_2، ...، $BS_{|B|}$ جملات موجود در متن B به‌ترتیبِ حضور در این متون و $|A|$ و $|B|$ تعداد جملات متون مورد نظر هستند.

مقایسهٔ جملات دو متن: هر جملهٔ AS_i از متن A با هر جملهٔ BS_j از متن B به‌صورت زیر مقایسه می‌شود:

تقسیم‌بندی جملات به لغات: جملات AS_i و BS_j به مجموعهٔ لغات تقسیم‌بندی می‌شود. فرض کنیم $\{AS_i^1, AS_i^2, ..., AS_i^n\}$ مجموعهٔ لغات جملهٔ AS_i و $\{BS_j^1, BS_j^2, ..., BS_j^m\}$ مجموعهٔ لغات جملهٔ BS_j باشند.

محاسبهٔ همانندی بین لغات جمله‌ای از متن A با جمله‌ای از متن B و برعکس: همانندی هر لغت AS_i^k از جملهٔ AS_i و هر لغت BS_j^l از جملهٔ BS_j که با $Sim_{AS_i^k \leftrightarrow BS_j^l}$ نمایش داده می‌شود، به‌صورت زیر محاسبه می‌شود:

◇ در صورتی که دو واژه یکسان باشند، $Sim_{AS_i^k \leftrightarrow BS_j^l}$ برابر با γ قرار می‌گیرد.

◇ در غیر این صورت، $Sim_{AS_i^k \leftrightarrow BS_j^l}$ به‌صورت زیر محاسبه می‌شود:

(۶)

$$Sim_{AS_i^k \leftrightarrow BS_j^l} = \frac{n_{AS_i^k BS_j^l}}{n_{AS_i^k} + n_{BS_j^l} - n_{AS_i^k BS_j^l}}$$

که $n_{AS_i^k}$ فراوانی لغت AS_i^k، $n_{BS_j^l}$ فراوانی لغت BS_j^l و $n_{AS_i^k BS_j^l}$ فراوانی هم‌رخدادی این دو لغت است که این فراوانی‌ها با توجه به پیکره‌ای از متون فارسی محاسبه شده‌اند.

همانندی هر لغت AS_i^k از جملهٔ AS_i با جملهٔ BS_j که با $Sim_{AS_i^k \leftrightarrow BS_j}$ نمایش داده می‌شود، برابر با بیشینهٔ $Sim_{AS_i^k \leftrightarrow BS_j^l}$ ($l = 1,...,m$) قرار داده می‌شود.

به همین ترتیب، همانندی هر واژهٔ BS_j^l از جملهٔ BS_j با جملهٔ AS_i محاسبه شده و با $Sim_{BS_j^l \leftrightarrow AS_i}$ نمایش داده می‌شود.

محاسبهٔ همانندی هر جمله از متن A با هر جمله از متن B بر اساس همانندی لغات آن‌ها: میزان همانندی جملهٔ AS_i به جملهٔ BS_j و میزان همانندی جملهٔ BS_j به AS_i به‌ترتیب با استفاده از فرمول‌های زیر محاسبه می‌شود:

(۷)

$$Sim_{AS_i \to BS_j} = \frac{\sum_{k=1}^{n} Sim_{AS_i^k \to BS_j}}{n}$$

و

(۸)

$$Sim_{BS_j \to AS_i} = \frac{\sum_{l=1}^{m} Sim_{BS_j^l \to AS_i}}{m}$$

تشکیل ماتریس همانندی: ابتدا ماتریس همانندی اولیه تشکیل می‌شود. سپس، با درنظرگرفتن اثر جملات همسایه، این ماتریس بهبود می‌یابد و ماتریس همانندی نهایی تشکیل می‌شود.

تشکیل ماتریس‌های اولیه همانندی: با استفاده از مقادیر محاسبه‌شده به‌عنوان میزان همانندی جملات مختلف دو متن، ماتریس‌های اولیهٔ $Sim_{A \to B} = \left[a_{ij}\right]_{i=1...|A|}^{j=1...|B|}$ و $Sim_{B \to A} = \left[b_{ij}\right]_{i=1...|A|}^{j=1...|B|}$ ساخته می‌شوند که

$$a_{ij} = \begin{cases} 2 \; if & Sim_{AS_i \to BS_j} > \alpha_1 \\ 1 \; if & \beta_1 < Sim_{AS_i \to BS_j} < \alpha_1 \\ 0 \; if & Sim_{AS_i \to BS_j} < \beta_1 \end{cases} \quad (۹)$$

و

$$b_{ij} = \begin{cases} 2 \; if & Sim_{BS_j \to AS_i} > \alpha_2 \\ 1 \; if & \beta_2 < Sim_{BS_j \to AS_i} < \alpha_2 \\ 0 \; if & Sim_{BS_j \to AS_i} < \beta_2 \end{cases} \quad (۱۰)$$

که α_1، α_2، β_1 و β_2 مقادیری آستانه‌ای بوده و با انجام آزمایشات محاسبه می‌شوند.

تشکیل ماتریس همانندی نهایی: با استفاده از مؤلفه‌های دو ماتریس به‌دست‌آمده در مرحلهٔ قبل، ماتریس نهایی همانندی $H = [h_{ij}]$ به‌صورت زیر ایجاد می‌شود:

$$h_{ij} = \begin{cases} 1 & if \; \left(a_{ij} \geq 1 \,\&\, b_{ij} \geq 1\right) \,\&\, \left(\exists k,l \in \{-1,0,1\} s.t.: \; (a_{i+k \; j+l} = 2 \,\&\, b_{i+k \; j+l} = 2)\right) \\ 0 & otherwise \end{cases} \quad (۱۱)$$

تشخیص جملات همانند: در نهایت، جملهٔ AS_i از متن A و جملهٔ BS_j از متن B همانند شناخته می‌شوند اگر $h_{ij}=1$.

پردازش نهایی: پس از مشخص شدن جملات مشابه دو متن، فرایند زیر بر روی متن مشکوک و متن منبع احتمالی انجام شده و قسمت‌های همانند تشخیص داده‌شدهٔ مجزا به یکدیگر متصل شده و قسمت‌های با طول کم حذف می‌شوند:

- ◇ در صورتی که دو جمله نزدیک به هم در متن مشکوک (در متن منبع احتمالی) همانند تشخیص داده شود و فاصلهٔ آن‌ها در متن کمتر از مقدار آستانه‌ای باشد، قسمت همانند تشخیص داده‌شده برابر با ابتدای جملهٔ اول تا انتهای جملهٔ دوم در نظر گرفته می‌شود. در پیاده‌سازی‌های صورت‌گرفته، از نصف طول دو جملهٔ همانند تشخیص داده‌شده به‌عنوان این مقدار آستانه‌ای استفاده شده است.
- ◇ در صورتی که قسمتی از متن همانند تشخیص داده شده و پس از انجام مرحلهٔ قبل طول آن همچنان کمتر از مقداری آستانه‌ای باشد، این قسمت از مجموعهٔ متون همانند تشخیص داده‌شده حذف می‌گردد. در پیاده‌سازی‌های صورت‌گرفته از ۴ کلمه به‌عنوان این مقدار آستانه‌ای استفاده شده است.

۳-۲. ارزیابی روش‌های همانندجوی پیشنهادی

در این بخش، ابتدا معیارهای ارزیابی یک روش همانندجو را بررسی کرده و در ادامه، پس از بیان جزئیات آزمایشات صورت‌گرفته بر روی روش‌های پیشنهادی، نتایج این آزمایشات را ارائه و روش‌های ارائه‌شده را با یکدیگر مقایسه می‌کنیم.

۳-۲-۱. معیارهای ارزیابی یک روش همانندجویی

در این بخش، معیارهای بررسی کیفیت یک روش همانندجویی ارائه می‌شود تا در بخش‌های بعدی با استفاده از آن‌ها، کیفیت الگوریتم‌های پیشنهادی بررسی گردد. برای ارزیابی کیفیت یک الگوریتم همانندجو از سه پارامتر دقت[1]، فراخوانی[2]، دانه‌دانه‌بودن[3] (Potthast et al. 2010) استفاده می‌شود. در ادامه، این پارامترها و نحوهٔ محاسبهٔ آن‌ها بیان خواهد شد.

فرض کنیم یک الگوریتم همانندجو را بر روی یک متن مشکوک و یک متن منبع احتمالی اجرا کنیم. مجموعهٔ R را به‌عنوان نمایشگر مجموعه موارد همانند تشخیص داده‌شده توسط الگوریتم در نظر می‌گیریم. همچنین، فرض می‌کنیم S نمایشگر مجموعه موارد همانند بین این دو متن باشد. در این صورت، پارامترهای دقت و فراخوان به‌صورت زیر تعریف می‌شوند:

$$prec(S,R)=\frac{1}{|R|}\sum_{r\in R}\frac{|\bigcup_{s\in S}s\bigcap r|}{|r|} \quad (۱۲)$$

و

$$rec(S,R)=\frac{1}{|S|}\sum_{s\in S}\frac{|\bigcup_{r\in R}s\bigcap r|}{|s|} \quad (۱۳)$$

که $s\bigcap r$ کاراکترهای همپوشان بین s و r است زمانی که آن‌ها در هر دو قسمت از متن مشکوک و متن منبع احتمالی حداقل یک کاراکتر مشترک داشته باشند. در غیر این صورت، $s\bigcap r$ تهی است (Stamatatos 2011).

در همانندجویی، پارامترهای دقت و فراخوان تصویری کلی از تأثیرگذاری روش‌ها ارائه نمی‌کنند. در شرایطی که یک الگوریتم همانندجو قسمت‌های همپوشان از متن منبع احتمالی را به‌عنوان منبع برای یک مورد همانندی ارائه کند، یا در مواردی که قسمت‌های

1. precision 2. recall 3. granularity

متن منبع احتمالی به قسمت‌های کوچک‌تر تقسیم شده و بازگردانده شوند، پارامترهای مورد نظر تغییر خواهند کرد (افزایش خواهند یافت). در نتیجه، به معیار دیگری نیاز است که این موارد را در نظر بگیرد. برای تعریف معیار جدید فرض کنیم $S_R \subseteq S$ موارد تشخیص داده‌شده در R باشند و $R_s \subseteq R$ موارد تشخیص داده‌شده مربوط به S باشند. حال، معیار دانه‌دانه‌بودن به‌صورت زیر تعریف می‌شود:

$$Gran(S,R) = \frac{1}{|S_R|}\sum_{s \in S_R} |R_S| \quad (۱۴)$$

کمترین مقدار برای این معیار ۱ است. هر قدر مقدار این پارامتر بیشتر باشد، قسمت‌های بیشتری از یک متن (احتمالاً همپوشان) در تناظر با یک مورد مشابه همانندی تشخیص داده شده است. پارامترهای دقت، فراخوانی و دانه‌دانه‌بودن را می‌توان با یکدیگر ترکیب و به‌صورت یک معیار کلی Plagdet که به‌صورت زیر تعریف می‌شود نمایش داد:

$$plag\ \det(S,R) = \frac{F_1}{\log_2(1+Gran(S,R))} \quad (۱۵)$$

که F_1 میانگین هارمونی[1] مقادیر پارامترهای دقت و فراخوان است. لازم به ذکر است که Plagdet پارامتر نهایی است که از آن در رتبه‌بندی الگوریتم‌های همانندجویی ارائه شده در مسابقات همانندجویی می‌شود.

۳-۲-۲. نتایج تجربی ارزیابی الگوریتم‌های همانندجویی پیشنهادی

در راستای بررسی نحوهٔ کارکرد الگوریتم‌های پیشنهادی، این الگوریتم‌ها بر روی مجموعه داده‌هایی پیاده‌سازی شده‌اند و آزمایش‌هایی بر اساس این روش‌ها انجام شده است که در ادامه، جزئیات فعالیت‌های انجام‌شده در این قسمت و نتایج آن بیان خواهد شد.

۳-۲-۳. نحوهٔ پیاده‌سازی

در این قسمت، ملاحظات صورت‌گرفته در پیاده‌سازی الگوریتم‌های پیشنهادی بیان می‌شود. همان‌طور که بیان شد، یکی از پیش‌پردازش‌های مورد نیاز در روش‌های پیشنهادی، انجام ریشه‌یابی لغوی است. در آزمایشات صورت‌گرفته، برای ریشه‌یابی

۱. میانگین هارمونیک (harmonic mean) چند پارامتر برابر است با تعداد آن‌ها تقسیم بر مجموع وارون آن‌ها.

لغوی از «ابزار پارسر زبان فارسی» (استیری و همکاران ۱۳۹۱) استفاده شده است. از دیگر پیش‌پردازش‌های مورد نیاز در الگوریتم‌های ارائه‌شده حذف ایست‌واژه‌هاست که در الگوریتم‌های پیشنهادی از لیست ایست‌واژه‌های فراهم‌شده در Taghva et al. (2003) استفاده شده است. ابزار دیگر مورد نیاز برای پیاده‌سازی الگوریتم معنایی پیشنهادی، وجود لغت‌نامه‌ای جهت بررسی نزدیکی معنایی دو واژه است. در پیاده‌سازی‌های صورت‌گرفته از «فرهنگ جامع واژگان مترادف و متضاد زبان فارسی» تألیف Khodaparasti (1997) استفاده شده است که دربرگیرندهٔ مترادف‌ها و متضادهای واژگان است.

الگوریتم همانندجوی احتمالاتی پیشنهادی نیز نیازمند محاسبهٔ فراوانی واژگان و فراوانی هم‌رخدادی واژگان با توجه به پیکره‌ای عظیم از متون فارسی است. برای محاسبهٔ این مقادیر و با توجه به قدرت محاسباتی در دسترس، از قسمتی از پیکرهٔ «همشهری» AleAhmad et al. (2009) ۲۰ درصد از داده‌های این پیکره که به‌صورت تصادفی انتخاب شده‌اند) به‌عنوان پیکرهٔ مرجع استفاده شده و جدول هم‌رخدادی لغات در هر دنباله ۳۱ کلمه‌ای پیاپی از متون این مجموعه محاسبه شده است.

پارامترهای مورد استفاده در آزمایشات صورت‌گرفته روی روش‌های پیشنهادی با استفاده از روش آزمون و خطا و با در نظر گرفتن مقادیر مختلف برای این پارامترها و انجام آزمایش روی قسمت کوچکی از مجموعهٔ دادهٔ آزمایش محاسبه شده‌اند. با توجه به تعداد زیاد پارامترهای دخیل در محاسبات، برای ممکن شدن محاسبهٔ مقادیر بهینهٔ تقریبی برای پارامترهای مورد نظر در روش معنایی پیشنهادی، از مجموعه‌ای از مفروضات استفاده شده که عبارت‌اند از:

- ◇ $\alpha_1 = \alpha_2$، $\alpha'_1 = \alpha'_2$، $\beta_1 = \beta_2$ و $\beta'_1 = \beta'_2$، و تنها مضارب صحیح ۰/۱ در محاسبات این پارامترها در نظر گرفته شده است.
- ◇ مقدار پارامتر γ مضربی صحیح از ۰/۱ و کمتر از ۱ است.
- ◇ δ عددی صحیح و بزرگ‌تر از ۱ است.

با توجه به این رویکرد، مقادیر بهینه برای پارامترهای α_1، α'_1، α_2، α'_2، β_1، β'_1، β_2، β'_2، γ، δ در الگوریتم همانندجوی معنایی ارائه‌شده به‌ترتیب برابر با ۰/۴، ۰/۴، ۰/۴، ۰/۴، ۰/۳، ۰/۳، ۰/۳، ۰/۳، ۰/۹، ۳ به‌دست آمده است.

به طریقی مشابه و با مفروضات زیر مقادیر بهینهٔ تقریبی برای پارامترهای α_1، α_2، β_1 و β_2 و γ در الگوریتم احتمالاتی پیشنهادی به‌ترتیب برابر با ۰/۰۳۵، ۰/۰۳۵، ۰/۰۲۵، ۰/۰۲۵و

۰/۱ در نظر گرفته شده است:

- ◇ $\alpha_1 = \alpha_2$، $\beta_1 = \beta_2$ و تنها مضارب صحیح ۰/۰۵ در محاسبات این پارامترها در نظر گرفته شده است.
- ◇ مقدار پارامتر γ مضربی صحیح از ۰/۰۱و کمتر از ۱ است.

لازم به ذکر است که در آزمایشات صورت‌گرفته تنها به‌صورت ساده و برای ممکن شدن آزمایشات سعی شده مقادیر بهینهٔ تقریبی برای پارامترهای مورد استفاده محاسبه شود. قطعاً با استفاده از روش‌های تنظیم پارامتر، می‌توان مقادیر دقیق بهینه برای پارامترهای مورد نظر را یافته و نتایج روش‌های پیشنهادی را بهبود داد. در اینجا اما، با توجه به عدم تمرکز این مقاله بر روی این مسئله، تنها به همین محاسبات بسنده کرده و محاسبات دقیق‌تر در این زمینه را به آینده موکول می‌کنیم.

۳-۲-٤. ارزیابی

برای ارزیابی الگوریتم‌های ارائه‌شده از مجموعه داده ارائه شده به‌عنوان داده آزمایش در کنفرانس تشخیص سرقت علمی در زبان فارسی[1] استفاده شده است. این مجموعه داده شامل مجموعه‌ای از متون منبع و مجموعه‌ای از متون مشکوک است که در آن متن منبع احتمالی هر متن مشکوک نیز مشخص شده است. علاوه بر این، قسمت‌های همانند موجود در هر زوج نیز مشخص شده که با استفاده از آن بررسی صحت نتایج الگوریتم‌های پیشنهادی ممکن خواهد بود.

موارد سرقت علمی موجود در این مجموعه به سه دسته تقسیم شده است:

۱. کپی دقیق شامل ۱۹۸ مورد سرقت علمی؛
۲. بازنویسی‌شده توسط کامپیوتر شامل ۱۶۰۷ مورد سرقت علمی؛
۳. بازنویسی‌شده توسط کاربر انسانی شامل ۱۴۷ مورد سرقت علمی.

برای ارزیابی الگوریتم‌های پیشنهادی، آن‌ها را بر روی زوج متون مشکوک و منبع احتمالی هر دسته اعمال کرده و نتایج آن‌ها را محاسبه می‌کنیم.

در دستهٔ «بازنویسی‌شده توسط کاربر انسانی» که شامل متون همانند شبیه‌سازی‌شده توسط کاربر انسانی است، مقادیر پارامترهای مورد نیاز برای بررسی کیفیت الگوریتم معنایی طراحی‌شده به شرح زیر محاسبه شده است:

1. http://www.ictrc.ac.ir/plagdet/

پارامتر دقت برابر با ۰/۸۸ و پارامتر فراخوان برابر با ۰/۸۴. با توجه به این مهم و مقدار یک به‌عنوان پارامتر دانه‌دانه‌بودن با توجه به پیوستگی قسمت‌های همانند تشخیص داده‌شده، پارامتر نهایی همانندجویی یعنی plagdet برابر با ۰/۸۵ می‌شود.

این آزمایش با داده‌های یکسان بر روی الگوریتم احتمالاتی پیشنهادی نیز اجرا شده و مقادیر پارامترهای مورد نظر به‌شرح زیر محاسبه شده است. پارامتر دقت برابر با ۰/۸۶ و پارامتر فراخوان برابر با ۰/۸۶. با توجه به این مهم و مقدار یک به‌عنوان پارامتر دانه‌دانه‌بودن، پارامتر نهایی همانندجویی یعنی plagdet برابر با ۰/۸۶ محاسبه شده است.

در دستهٔ «بازنویسی‌شده توسط کامپیوتر» که شامل متون همانند شبیه‌سازی‌شده توسط کامپیوتر است، مقادیر پارامترهای مورد نیاز برای بررسی کیفیت الگوریتم معنایی طراحی‌شده به شرح زیر محاسبه شده است: پارامتر دقت برابر با ۰/۹۰ و پارامتر فراخوان برابر با ۰/۹۱. با توجه به این مهم و مقدار یک به‌عنوان پارامتر دانه‌دانه‌بودن با توجه به پیوستگی قسمت‌های همانند تشخیص داده‌شده، پارامتر نهایی همانندجویی یعنی plagdet برابر با ۰/۹۰ می‌شود.

این آزمایش با داده‌های یکسان بر روی الگوریتم احتمالاتی پیشنهادی نیز اجرا شده و مقادیر پارامترهای مورد نظر به شرح زیر محاسبه شده است. پارامتر دقت برابر با ۰/۸۹ و پارامتر فراخوان برابر با ۰/۹۴. با توجه به این مهم و مقدار یک به‌عنوان پارامتر دانه‌دانه‌بودن با توجه به پیوستگی قسمت‌های همانند تشخیص داده‌شده، پارامتر نهایی همانندجویی یعنی plagdet برابر با ۰/۹۱ می‌شود.

در دستهٔ «کپی دقیق» که شامل متون دقیقاً کپی شده است، مقادیر پارامترهای مورد نیاز برای بررسی کیفیت الگوریتم معنایی طراحی‌شده به شرح زیر محاسبه شده است: پارامتر دقت برابر با ۰/۹۲ و پارامتر فراخوان برابر با ۰/۹۸. با توجه به این مهم و مقدار یک به‌عنوان پارامتر دانه‌دانه‌بودن با توجه به پیوستگی قسمت‌های همانند تشخیص داده‌شده، پارامتر نهایی همانندجویی یعنی plagdet برابر با ۰/۹۴ می‌شود.

این آزمایش با داده‌های یکسان بر روی الگوریتم احتمالاتی پیشنهادی نیز اجرا شده و مقادیر پارامترهای مورد نظر به شرح زیر محاسبه شده است. پارامتر دقت برابر با ۰/۹۰ و پارامتر فراخوان برابر با ۰/۹۸. با توجه به این مهم و مقدار یک به‌عنوان پارامتر دانه‌دانه‌بودن، پارامتر نهایی همانندجویی یعنی plagdet برابر با ۰/۹۳ می‌شود.

جدول ۱، به‌طور خلاصه، عملکرد دو الگوریتم پیشنهادی در این پژوهش را در

دسته‌های مختلف متون آزمایشی نشان می‌دهد. نتایج گزارش‌شده نشان‌دهندهٔ این است که در حالی‌که مقدار پارامتر دقت برای روش معنایی پیشنهادی به‌طور متوسط حدود ۲ درصد بالاتر است، مقدار پارامتر فراخوانی این روش به‌طور متوسط حدود ۲ درصد کمتر از مقدار این پارامتر برای روش احتمالاتی پیشنهادی است.

جدول ۱. مقایسهٔ کیفیت الگوریتم‌های پیشنهادی

الگوریتم	کپی دقیق				بازنویسی‌شده توسط کامپیوتر				بازنویسی‌شده توسط کاربر انسانی			
	دقت	فراخوانی	دانه‌دانه‌بودن	plagdet	دقت	فراخوانی	دانه‌دانه‌بودن	plagdet	دقت	فراخوانی	دانه‌دانه‌بودن	plagdet
معنایی	۰/۹۲	۰/۹۸	۱	۰/۹۴	۰/۹۰	۰/۹۰	۱	۰/۹۰	۰/۸۸	۰/۸۴	۱	۰/۸۵
احتمالاتی	۰/۹۰	۰/۹۸	۱	۰/۹۳	۰/۸۹	۰/۹۴	۱	۰/۹۱	۰/۸۶	۰/۸۶	۱	۰/۸۶

همان‌طور که بیان شد، یکی از نوآوری‌های روش‌های ارائه‌شده (حتی نسبت به روش‌های ارائه‌شده برای سایر زبان‌ها) این است که در بررسی همانندی دو جمله، همانندی جملات همسایهٔ آن‌ها نیز در نظر گرفته می‌شود. برای بررسی میزان اثربخشی این اقدام در همانندجویی، این مرحله را (مرحلهٔ تشکیل ماتریس همانندی نهایی از روش‌های معنایی و احتمالاتی که در آن ماتریس نهایی محاسبه می‌شود) از روش‌های ارائه‌شده حذف کرده و نتایج همانندجویی با روش‌های تغییریافته را بررسی می‌کنیم. این نتایج در جدول ۲، گزارش شده است. همان‌طور که مشاهده می‌شود، حذف مراحل ذکرشده در الگوریتم‌های ارائه‌شده تأثیر چندانی در پارامترهای دقت و فراخوانی همانندجویی در دستهٔ کپی دقیق از موارد سرقت علمی نداشته، لیکن در دسته‌های بازنویسی‌شده، علی‌رغم ثابت ماندن تقریبی پارامتر دقت، پارامتر فراخوانی به‌طور محسوس (حدود ۵ درصد) کاهش یافته است. به‌عبارت دیگر، با در نظر گرفتن همانندی جملات همسایه دو جمله در بررسی همانندی آن‌ها، قدرت تشخیص همانندی متون بازنویسی‌شده بر اساس روش‌های پیشنهادی به میزان ۵ درصد افزایش یافته است.

جدول ۲. مقایسهٔ کیفیت الگوریتم‌های پیشنهادی بدون در نظر گرفتن تأثیر همانندی دو جمله بر روی همانندی جملات همسایهٔ آن‌ها

الگوریتم	کپی دقیق			بازنویسی‌شده توسط کامپیوتر				بازنویسی‌شده توسط کاربر انسانی			
	دقت	فراخوانی	دانه‌دانه‌بودن	plagdet	دقت	فراخوانی	دانه‌دانه‌بودن	plagdet	دقت	فراخوانی	دانه‌دانه‌بودن	plagdet
معنایی	۰/۹۳	۰/۹۸	۱	۰/۹۵	۰/۹۱	۰/۸۷	۱	۰/۸۸	۰/۸۸	۰/۸۰	۱	۰/۸۳
احتمالاتی	۰/۹۱	۰/۹۸	۱	۰/۹۴	۰/۸۹	۰/۹۰	۱	۰/۸۹	۰/۸۷	۰/۸۰	۱	۰/۸۳

علاوه بر این، برای مقایسهٔ زمان اجرای روش‌های همانندجویی پیشنهادی زمان پردازش لازم برای بررسی همانندی تمام موارد موجود در هر دسته از موارد سرقت علمی را محاسبه کرده و در جدول ۳، با یکدیگر مقایسه می‌کنیم. لازم به ذکر است که آزمایشات صورت‌گرفته بر روی کامپیوتری با پردازندهٔ مرکزی "3.7 GHz Intel i3"، حافظهٔ داخلی "4 GB" و سیستم عامل ۶۴ بیتی ویندوز ۷ انجام شده است. علاوه بر این، با توجه به یکسانی پیش‌پردازش‌های مورد نیاز در هر دو روش پیشنهادی، زمان‌های اجرای بیان‌شده بدون در نظر گرفتن زمان مورد نیاز برای پیش‌پردازش متون گزارش شده‌اند. نتایج محاسبات ارائه‌شده در آخرین ستون جدول ۳، بیانگر آن است که روش احتمالاتی پیشنهادی بسیار کاراتر بوده و میانگین زمان صرف‌شده برای بررسی هر مورد سرقت علمی با استفاده از آن تنها ۳/۸ درصد زمان صرف‌شده توسط روش معنایی پیشنهادی است. بنابراین، روش همانندجویی احتمالاتی پیشنهادی از نظر سرعت پردازش بسیار سریع‌تر از روش همانندجویی معنایی عمل می‌کند.

جدول ۳. مقایسهٔ زمان اجرای الگوریتم‌های پیشنهادی

الگوریتم	کپی دقیق	بازنویسی‌شده توسط کامپیوتر	بازنویسی‌شده توسط کاربر انسانی	میانگین زمان صرف‌شده برای بررسی هر مورد سرقت علمی
معنایی	۵۳۰ ثانیه	۴۲۸۰ ثانیه	۳۳۷ ثانیه	۲/۶۳۶ ثانیه
احتمالاتی	۱۸ ثانیه	۱۶۷ ثانیه	۱۴ ثانیه	۰/۱۰۱ ثانیه

در خاتمه، لازم به ذکر است که متأسفانه با توجه به عدم دسترسی به پیاده‌سازی سایر روش‌های ارائه‌شده برای همانندجویی در زبان فارسی، عدم دسترسی به مجموعه دادهٔ مورد استفاده برای آزمایش این روش‌ها و همچنین، فقدان نتایج ارزیابی این روش‌ها روی

مجموعه دادهٔ مورد بررسی در این مقاله، امکان مقایسهٔ روش‌های پیشنهادی با این روش‌ها ممکن نیست.

٤. نتیجه‌گیری

در این مقاله دو روش همانندجویی جدید _روش معنایی و روش احتمالاتی_ با هدف تشخیص متون بازنویسی‌شده در زبان فارسی ارائه شده است. در هر دو روش پیشنهادی، ابتدا پیش‌پردازش‌های مختص زبان فارسی بر روی متون مورد آزمایش انجام می‌شود. در روش همانندجویی معنایی پیشنهادی از ابزار لغت‌نامه برای محاسبهٔ همانندی جملات بازنویسی‌شده استفاده شده است. در روش همانندجویی احتمالاتی، همانندی واژگان بر اساس تابع احتمالی که با اتکا به هم‌رخدادی آن‌ها تعریف شده، محاسبه شده است. روش احتمالاتی پیشنهادی اولین روش همانندجویی احتمالاتی ارائه‌شده برای زبان فارسی بوده و روش معنایی پیشنهادی در مقایسه با روش‌های معنایی موجود از معیارهای جدیدی برای بررسی همانندی متون استفاده می‌کند. علاوه بر این، در حالی که در سایر روش‌های موجود همانندیِ هر جمله از متن مشکوک با هر جمله از متن منبع احتمالی به‌صورت مستقل بررسی می‌شود، در روش‌های پیشنهادی همانندی جملات همسایه نیز در بررسی همانندی دو جمله در نظر گرفته شده است. نتایج پیاده‌سازی و آزمایشات صورت‌گرفته بر روی روش‌های پیشنهادی در حالی که بیانگر کیفیت مناسب و تقریباً یکسان هر دو روش ارائه‌شده است، نشان می‌دهد که روش احتمالاتی پیشنهادی از کارایی بسیار بهتری برخوردار است و در زمان بسیار کمتری می‌تواند نتیجهٔ همانندجویی را ارائه نماید.

فهرست منابع

استیری، احمد، محسن کاهانی، رضا سعیدی، و احسان عسگریان. ۱۳۹۱. طراحی ابزار پارسر زبان فارسی. *اولین کنفرانس پردازش خط و زبان فارسی*. سمنان.

References

Adam, A. R. 2014. Plagiarism detection algorithm using natural language procesing based on grammar analyzing. *Journal of Theoretical & Applied Information Technology* 63 (1): 168-178.

AleAhmad, A., H. Amiri, E. Darrudi, M. Rahgozar, and F. Oroumchian. 2009. Hamshahri: A Standard Persian Text Collection. *Knowledge-Based Systems* 22 (5): 382–387.

Alzahrani, S. M., and N. Salim2009 .. On the use of fuzzy information retrieval for gauging similarity of Arabic documents. *Second International Conference on Applications of Digital Information and Web*

Technologies, 2009. ICADIWT'09. Second International Conference on the. IEEE: 539-544.

Alzahrani, S. M., N. Salim, and A. Abraham2012 .. Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems*, Man, and Cybernetics, Part C (Applications and Reviews) 42 (2): 133-149.

Alzahrani, S. M., N. Salim, and V. Palade2015 .. Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. *Journal of King Saud University-Computer and Information Sciences* 27 (3): 248-268.

Gharavi, E., K. Bijari, K. Zahirnia, and H. Veisi. 2016. A Deep Learning Approach to Persian Plagiarism Detection. *Working notes of FIRE 2016-Forum for Information Retrieval Evaluation*. Tehran, Iran. pp.:154-159.

Gipp, B., and N. Meuschke. 2011. Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence. *Proceedings of the 11th ACM symposium on Document engineering*: 249-258.

Khodaparasti, F. 1997. *A Comprehensive Dictionary of Persian Synonyms and Antonyms*. Shiraz: Danesh-Nameh Farsi Publication.

Koberstein, J., and Y.-K. Ng. 2006. Using word clusters to detect similar web documents. *International Conference on Knowledge Science, Engineering and Management*. Guilin City, China. pp. 215-228.

Le, Q. and T. Mikolov. 2014. Distributed representations of sentences and documents. *In International Conference on Machine Learning* Beijing, China. pp. 1188-1196.

Li, Y., D. McLean, Z. A. Bandar, J. D. O'shea, and K. Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE transactions on knowledge and data engineering* 18 (8): 1138-1150.

Mahdavi, P., Z. Siadati, and F. Yaghmaee. 2014. Automatic external Persian plagiarism detection using vector space model. *In Computer and Knowledge Engineering (ICCKE), 2014 4th International eConference on Computer and Knowledge Engineering*. Mashhad, Iran. pp. 697-702.

Mahmoodi, M. and M. M. Varnamkhasti. 2014. Design a Persian Automated Plagiarism Detector (AMZPPD). *International Journal of Engineering Trends and Technology* 8 (8): 465-467.

Mashhadirajab, F., and M. Shamsfard. 2016. A Text Alignment Algorithm Based on Prediction of Obfuscation Types Using SVM Neural Network. *Working notes of FIRE 2016-Forum for Information Retrieval Evaluation*. Tehran, Iran. pp.: 167-171.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. *In Advances in neural information processing systems.* Lake Tahoe. pp.: 3111-3119.

Minaei, B., and M. Niknam. 2016. An n-gram based Method for Nearly Copy Detection in Plagiarism Systems. *Working notes of FIRE 2016-Forum for Information Retrieval Evaluation*. Tehran, Iran. pp.: 172-175.

Potthast, M., B. Stein, A. Barrón-Cedeño, and P. Rosso. 2010. An evaluation framework for plagiarism detection. *In Proceedings of the 23rd international conference on computational linguistics: Posters*. Association for Computational Linguistics. Posters. Beijing, China. pp. 997-1005. Association for Computational Linguistics.

Potthast, M., M. Hagen, T. Gollub, M. Tippmann, J. Kiesel, P. Rosso, E. Stamatatos, and B. Stein. 2013. *Overview of the 5th International Competition on Plagiarism Detection*. In: Forner, P., Navigli, R., Tufis, D. (eds.) Working Notes Papers of the CLEF 2013 Evaluation Labs.

Rafieian, S. 2016. Plagiarism checker for Persian (PCP) texts using hash-based tree representative fingerprinting. *Journal of AI and Data Mining* 4 (2): 125-133.

Stamatatos, E. 2011. Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology* 62: 2512-2527.

Taghva, K., R. Beckley, and M. Sadeh. 2003. A list of Farsi stopwords. *Technical Report 2003-01, Information Science Research Institute, University of Nevada, Las Vegas*, July 2003.

Yerra, R., and Y.-K. Ng. 2005. A sentence-based copy detection approach for web documents. *International Conference on Fuzzy Systems and Knowledge Discovery*. Changsha, China. pp.: 481-482.

نصراله پاک‌نیت

متولد سال ۱۳۶۵، دارای مدرک دکتری در رشتهٔ ریاضی از دانشگاه شهید بهشتی تهران است. ایشان هم‌اکنون استادیار پژوهشکدهٔ علوم اطلاعات پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. رمزنگاری، الگوریتم‌ها و متن‌کاوی از جمله علایق پژوهشی وی است.

آزاده محبی

متولد سال ۱۳۵۷، دارای مدرک دکتری در رشتهٔ مهندسی طراحی سیستم‌ها از دانشگاه واترلو کانادا است. ایشان هم‌اکنون استادیار پژوهشکدهٔ فناوری اطلاعات پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است.

داده‌کاوی، سیستم‌های هوشمند، بازشناسی الگو، متن‌کاوی و بازیابی اطلاعات از جمله علایق پژوهشی وی است.