آزمایشگاه شبکه‌های اطلاعاتی و کامپیوتری
دانشکده علوم و فنون نوین، دانشگاه تهران

تشخیص سرقت علمی با رویکرد مبتنی بر گراف

ارائه دهنده: مژگان ممتاز(m.momtaz92@ut.ac.ir)

ارائه مربوط به همایش بررسی راهکارهای پیشگیری از سرقت علمی

سایر نویسندگان: دکتر مصطفی صالحی، دکتر هادی ویسی، کیوان بیجاری

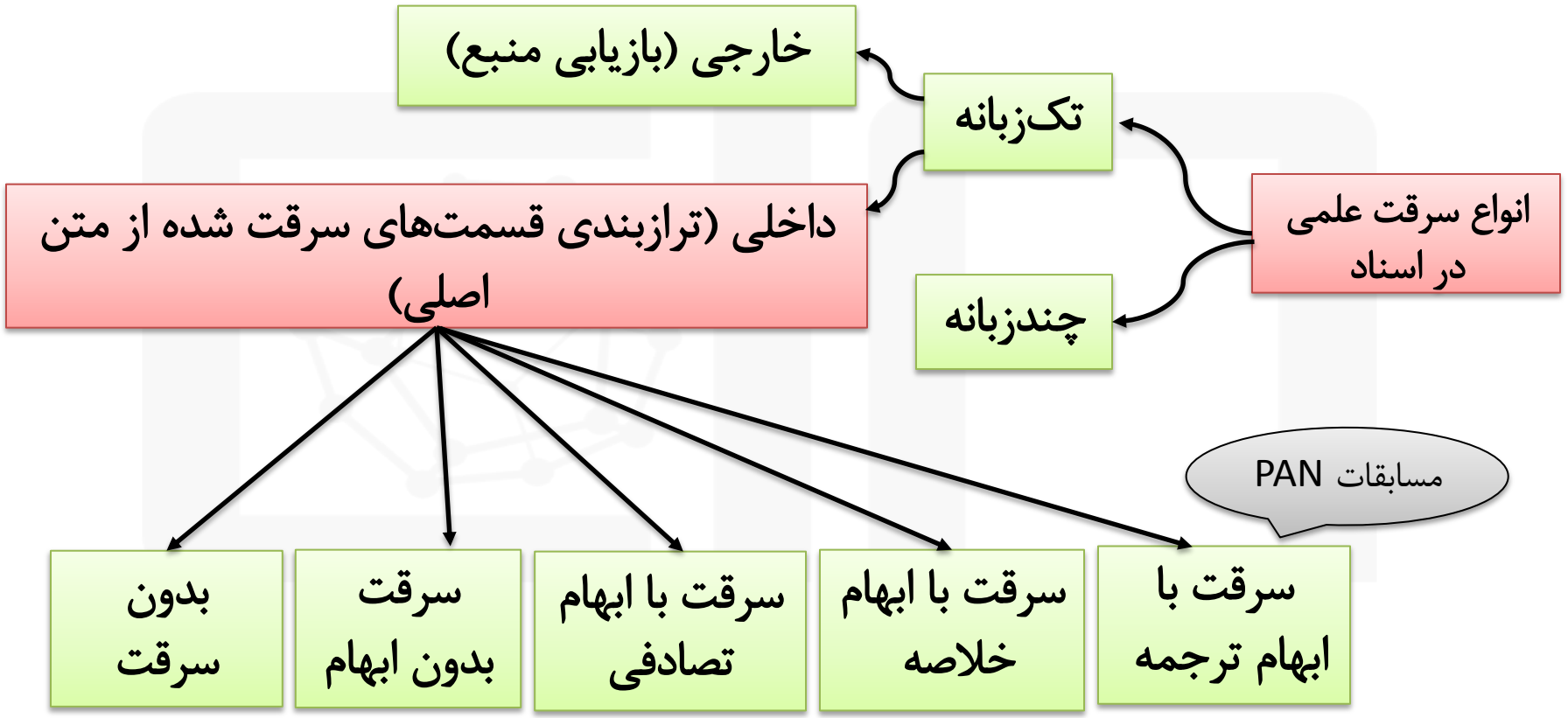
بهمن ۹۵

فهرست مطالب

تعریف موضوع–سرقت علمی

سرقت علمی- ادبی استفاده از ایده‌ها و عبارات دیگران، به عنوان ایده و عبارات خویش و در واقع نوعی دستبرد فکری یا ایده دزدی است. سرقت علمی می‌تواند شامل طیف وسیعی از دستبردهای آگاهانه تا کپی کردن اتفاقی مطالب دیگران باشد.

تعریف موضوع–انواع سرقت علمی در اسناد

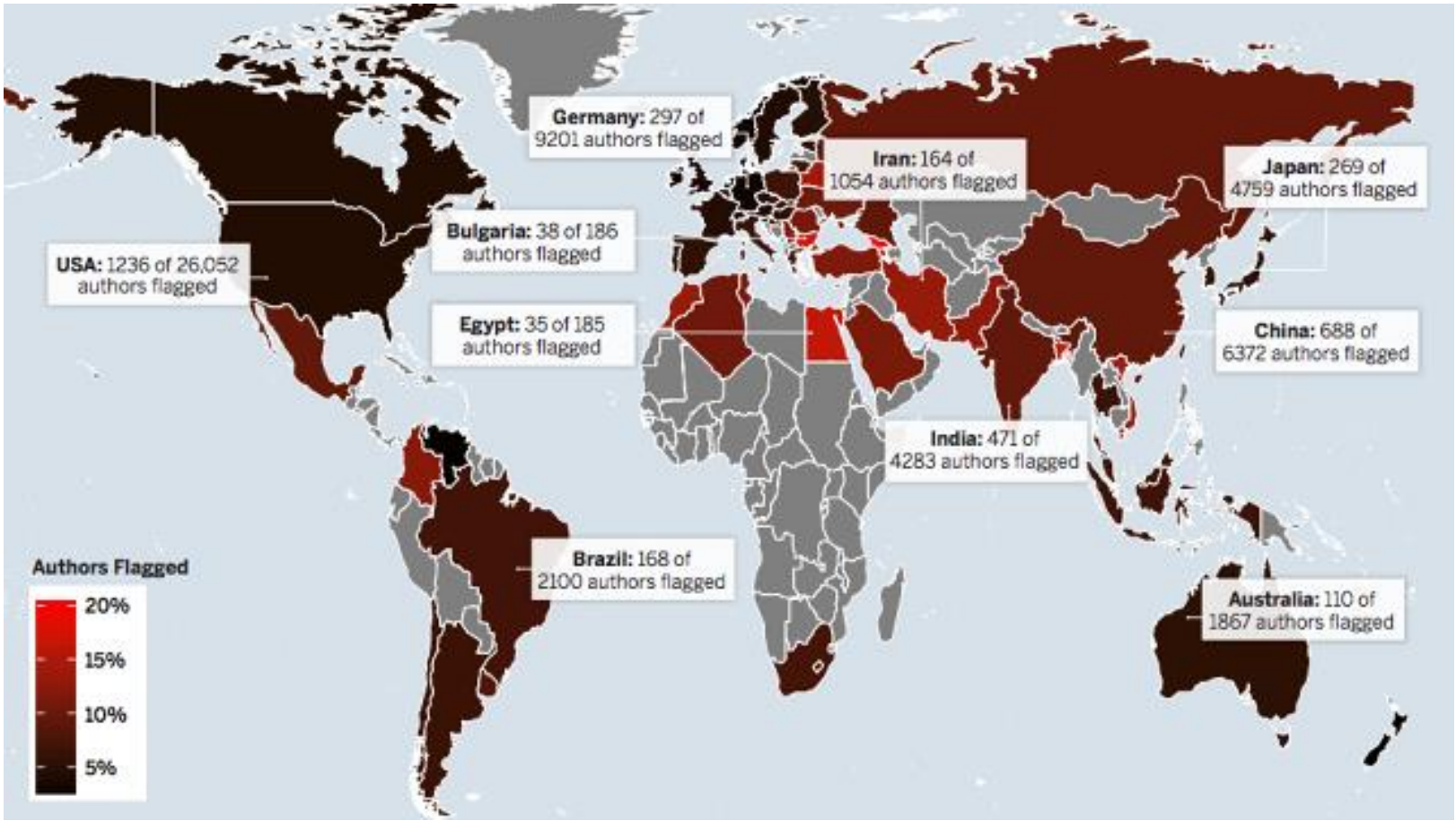


اهمیت موضوع

- بر اساس گزارش وبسایت نشریه Sience، همه کشورهای مورد مطالعه کمابیش درگیر مسئله سرقت علمی در مجامع علمی هستند ولی این پدیده در برخی از کشورها بیشتر رخ داده است.

- در این تحقیق گزارش شده است که بیش از پانزده درصد پژوهشگران ایرانی که در مقالات خود سرقت علمی داشته‌اند.

اهمیت موضوع–۱



شکل ۱- نقشه سرقت علمی در جهان (sciencemag.org)

روش‌های پیشین تشخیص سرقت علمی

شماره	روش
۱	پیدا کردن مشابهت به کمک روش انگشت‌نگاری
۲	پیدا کردن مشابهت با روش n -گرام
۳	روش مبتنی بر خوشه‌بندی (بازیابی منبع)
۴	پیدا کردن مشابهت با روش ساختاری
۵	روش‌های مبتنی بر شباهت معنایی
۶	روش مبتنی بر گراف

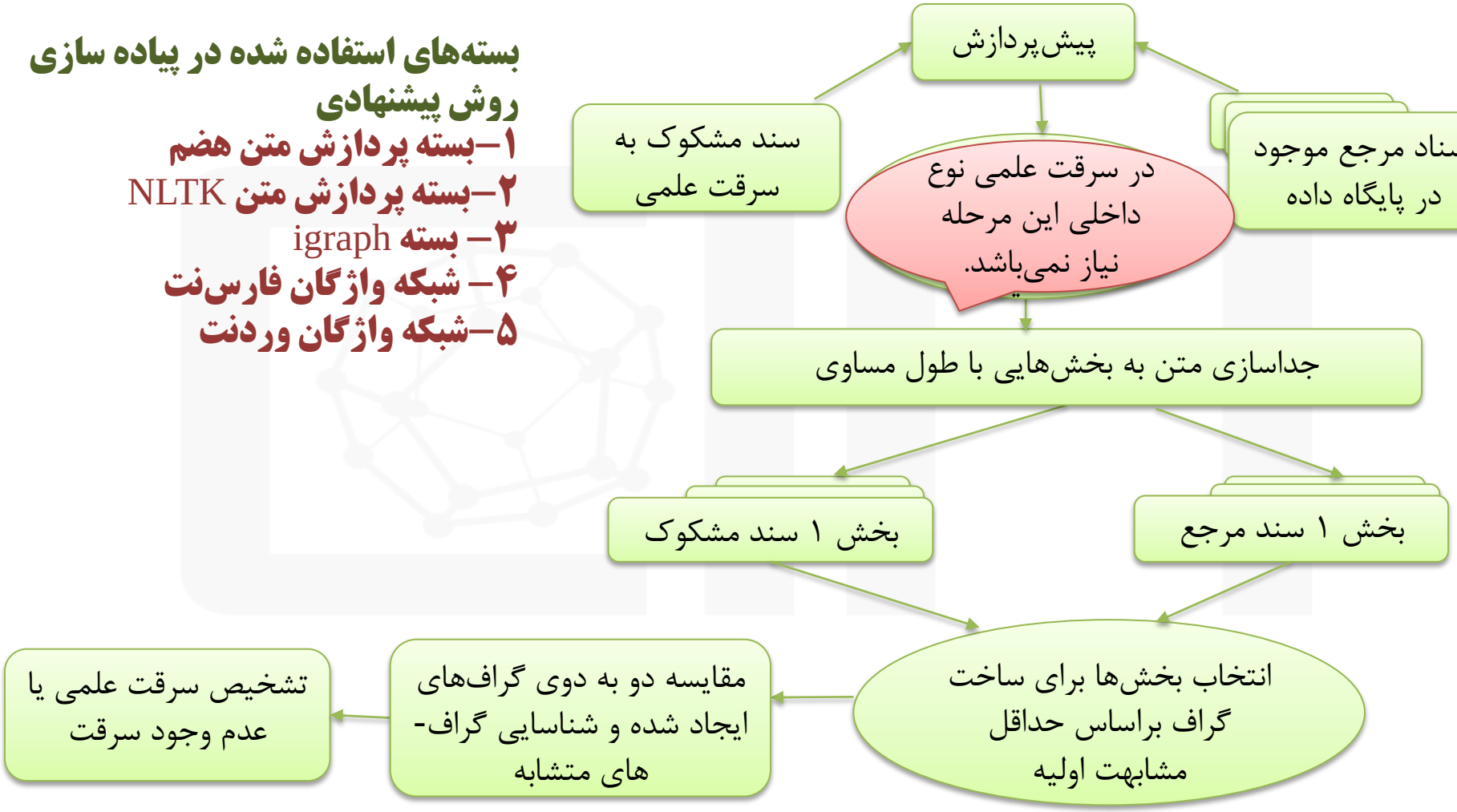
مشکلات روش‌های پیشین نمایش متن

مشکلات روش‌های پیشین	رفع مشکلات در روش پیشنهادی
بی‌توجهی به ترتیب کلمات	استفاده از گراف جهت‌دار
بی‌توجهی به ساختار متن	توجه به کل جمله
بی توجهی به کلمات مترادف استفاده شده	امکان اضافه کردن کلمات مترادف به گراف متناظر متن
دقت پایین در تشخیص سرقت علمی از نوع خلاصه	تشخیص بهتر سرقت علمی نوع خلاصه با توجه به ویژگی انتخاب تعداد گره‌های همسایه‌های بیشتر

متدلوژی و روش– نمودار مراحل تشخیص سرقت علمی سطح بازیابی منبع

بسته‌های استفاده شده در پیاده سازی روش پیشنهادی

۱–بسته پردازش متن هضم
۲–بسته پردازش متن NLTK
۳– بسته igraph
۴– شبکه واژگان فارس‌نت
۵–شبکه واژگان وردنت



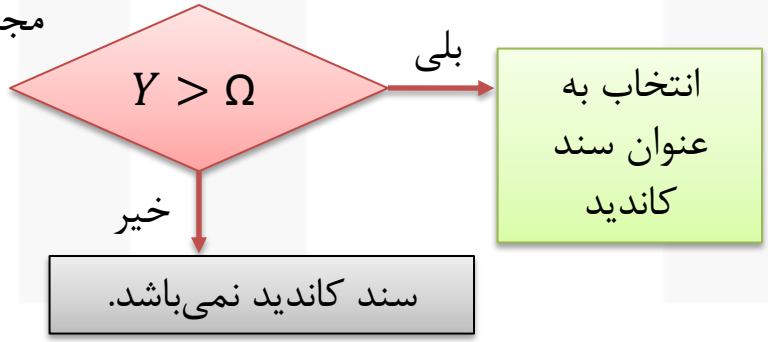
متدلوژی و روش–۱

مرحله پیش‌پردازش

- استفاده از بسته هضم در نرمال‌سازی متن
- برچسب‌زنی جز کلام (برای مرحله انتخاب سندهای کاندید)

مرحله انتخاب سندهای کاندید (سرقت علمی نوع خارجی)

- S1: مجموعه دوتایی‌ها سند مشکوک به سرقت علمی
- S2: مجموعه دوتایی‌ها سند مرجع
- تعداد دوتایی‌های مشترک=$Y=|S1 \cap S2|$
- آستانه تعداد دوتایی‌های مشترک $= \Omega$



- نکته: دوتایی‌ها تنها شامل کلمات با تگ اسم می‌باشند و سایر تگ‌ها مانند صفت‌ها و قیدها حذف شده‌اند.

متدلوژی و روش–۲

مرحله جداسازی متن

- جداسازی متن به مجموعه‌ای از پاراگراف‌ها ----------تشخیص سرقت علمی با ابهام خلاصه
- جداسازی متن به مجموعه‌ای از جملات---------- تشخیص سرقت علمی سایر سطوح داخلی
- جداسازی متن به بخش‌هایی با طول مشخصی از کاراکترها
- جداسازی متن به بخش‌هایی با تعداد مشخصی از کلمات-------تشخیص سرقت علمی سطح خارجی

معیار مشابهت اولیه جملات

- شباهت کسینوسی

متدلوژی و روش–۳

مرحله ساخت گراف برای بخش‌هایی با حداقل مشابهت اولیه

گره: سند، مفهوم، پاراگراف،جمله، گروه اسمی،کلمه

برقراری یال: دستوری، معنایی، آماری، رخداد همزمان کلمات

- ✓ **جمله**
- ✓ **پاراگراف**
- ✓ **متن**
- ✓ **پنجره با اندازه ثابت**

در روش پیشنهادی:

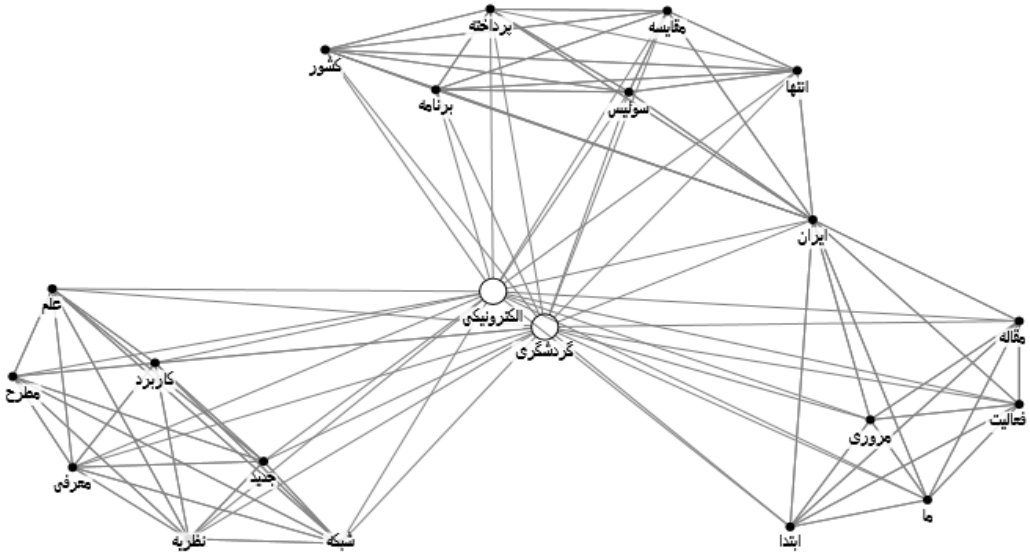
گره: کلمات (به جز ایست‌واژه‌ها)

برقراری یال: رخداد همزمان کلمات در پنجره با اندازه ثابت

متدلوژی و روش–۴

مرحله ساخت گراف برای بخش‌هایی با حداقل مشابهت اولیه–ادامه

- نمونه‌ای از تبدیل متن به گراف
- در این مقاله ما درابتدا به مروری بر فعالیت‌های گردشگری الکترونیکی در ایران پرداخته‌ایم و سپس مقایسه‌ای بین برنامه‌های گردشگری الکترونیکی ایران با کشور سوئیس ارائه شده است، و در انتها پس از معرفی علم جدید نظریه شبکه‌ها، کاربرد آن را در گردشگری الکترونیکی مطرح می کنیم.



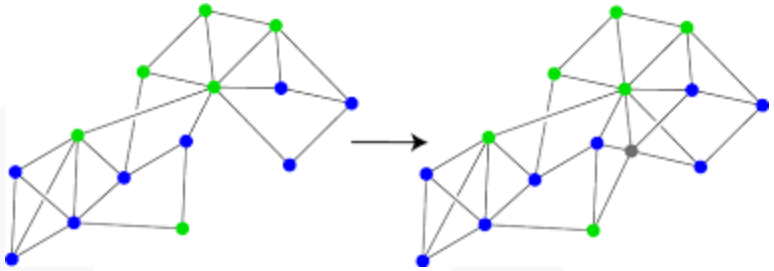
- ✓ رخداد همزمان کلمات در جملات

متدلوژی و روش–۵

مرحله تشخیص گراف‌های مشابه

روش تکراری

- ✓ روش‌های تکراری
- ✓ استخراج ویژگی
- ✓ فاصله ویرایشی



- ⊛ A: همسایگان درجه اول گره مشترک گراف سند مشکوک
- ⊛ B: همسایگان درجه اول گره مشترک گراف سند مرجع
- ⊛ $similarity(A,B) = \frac{A \cap B}{max(len(A),len(B))}$

α: آستانه مشابهت گره‌ها

if $(similarity(A,B)) > \alpha$ then A similar to B

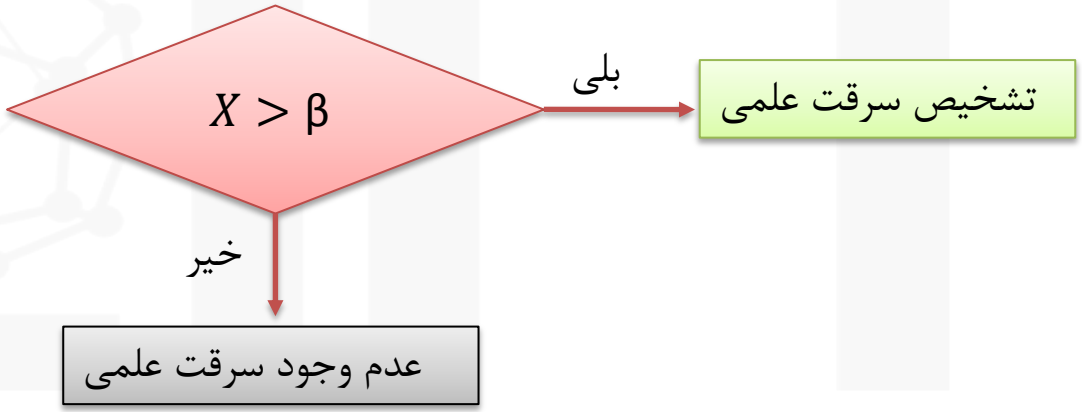
متدلوژی و روش–۶

مرحله تشخیص گراف‌های مشابه–ادامه

آستانه تعداد گره‌ها مشابه مشخص می‌کند آیا دو گراف مشابه می‌باشند یا خیر؟

X:تعداد گره‌های مشابه

β:آستانه تعداد گره‌ها مشابه



ارزیابی– مجموعه داده فارسی(بازیابی منبع)

- مجموعه داده سرقت علمی معنایی و با درجه ابهام بالا

تعداد سند مشکوک	تعداد سند مرجع
۱۱۲	۱۵۰۰

- مجموعه داده همه سطوح سرقت علمی (معنایی، با ابهام، بدون ابهام، بدون سرقت علمی)

تعداد سند مشکوک	تعداد سند مرجع
۳۰۰	۱۵۰۰

ارزیابی– مجموعه داده فارسی (ترازبندی متن)

سطح سرقت علمی	تعداد جفت سند
بدون سرقت علمی	۶۵۲
سرقت علمی بدون ابهام	۱۲۹
سرقت علمی با ابهام تصادفی	۲۸۲

ارزیابی– مجموعه داده انگلیسی(ترازبندی متن)

مجموعه داده آموزش و آزمایش

سطح سرقت علمی	تعداد جفت سند
بدون سرقت علمی	۱۰۰۰
سرقت علمی بدون ابهام	۱۰۰۰
سرقت علمی با ابهام تصادفی	۱۰۰۰
سرقت علمی با ابهام ترجمه	۱۰۰۰
سرقت علمی با ابهام خلاصه	۱۱۸۵

معیارهای ارزیابی- بازیابی منبع

$$\text{دقت} = \frac{\text{تعداد سندهای درست بازیابی شده}}{\text{تعداد سندهای اشتباه بازیابی شده+تعداد سندهای درست بازیابی شده}}$$

$$\text{فراخوانی} = \frac{\text{تعداد سندهای بازیابی شده}}{\text{تعداد سندهای بازیابی نشده+تعداد سندهای بازیابی شده}}$$

$$\text{معیار F} = 2 * \frac{\text{فراخوانی}*\text{دقت}}{\text{فراخوانی}+\text{دقت}}$$

معیارهای ارزیابی- ترازبندی متن

$$\text{دقت} = \frac{\text{تعداد بخش‌های درست بازیابی شده}}{\text{تعداد بخش‌های درست بازیابی شده+تعداد بخش‌های اشتباه بازیابی شده}}$$

$$\text{فراخوانی} = \frac{\text{تعدادبخش‌های بازیابی شده}}{\text{تعداد بخش‌های بازیابی نشده+تعداد بخش‌های بازیابی شده}}$$

$$\text{معیار F} = 2 * \frac{\text{فراخوانی*دقت}}{\text{فراخوانی+دقت}}$$

کد ارزیابی در سایت مسابقات PAN قرار داده شده است.

- نتایج بدست آمده برای تشخیص سرقت علمی اسناد فارسی
  - ۱- تاثیر پارامتر Ω
  - ۲- سطح بازیابی منبع
  - ۳- سطح ترازبندی متن

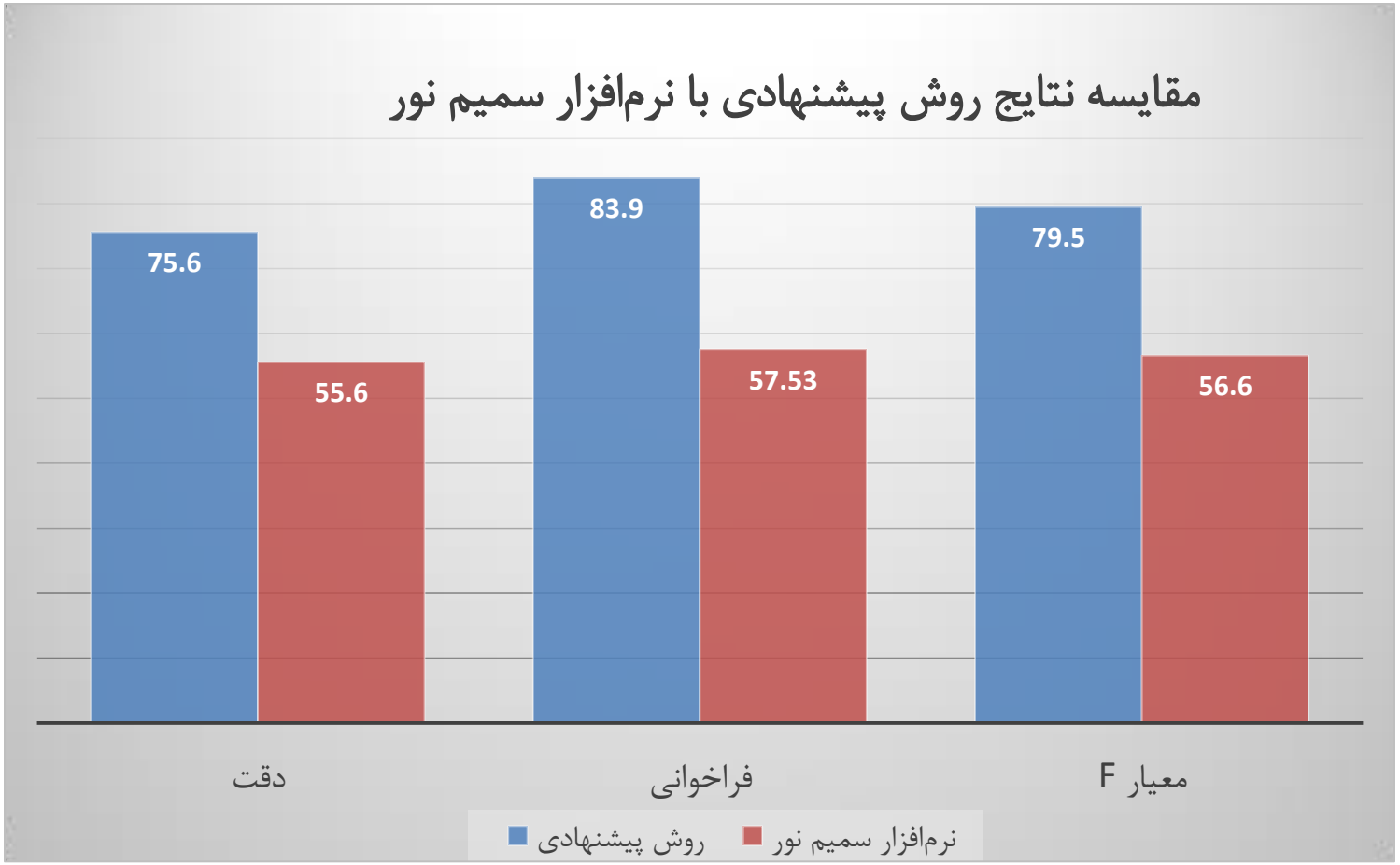
Ωارزیابی– تاثیر پارامتر

(تعداد دوتایی‌های مشترک) متفاوت در مجموعه Ωکارایی روش پیشنهادی براساس مقادیر مختلف داده سرقت معنایی و با درجه ابهام بالا

Ωحدآستانه	دقت	فراخوانی	معیار F
Ω =۲۰	۸۹.۷۰	۴۸.۷۵	۶۳.۱۷
Ω =۱۵	۸۸.۹۵	۵۵.۶۷	۶۸.۴۸
Ω =۱۰	۸۴.۹۱	۶۶.۹۶	۷۴.۸۷
Ω =۶	۷۵.۶	۸۳.۹	۷۹.۵

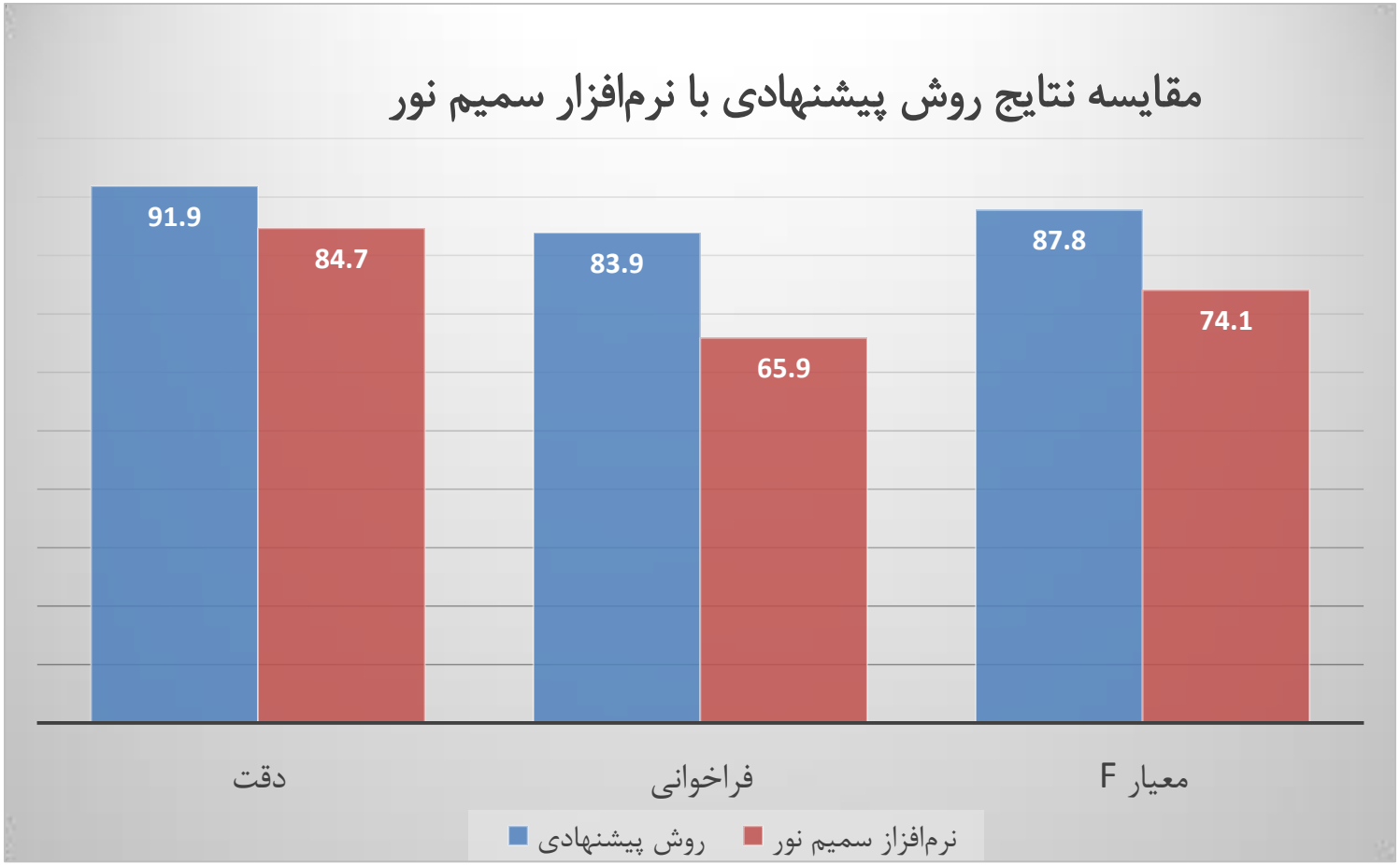
ارزیابی– نتایج بخش بازیابی سند اسناد فارسی

- مجموعه داده سرقت معنایی و با درجه ابهام بالا



ارزیابی– نتایج بخش بازیابی سند اسناد فارسی

- مجموعه داده همه سطوح سرقت علمی



ارزیابی– نتایج بخش ترازبندی سند اسناد فارسی

	دقت	فراخوانی	معیار F
سرقت علمی بدون ابهام	۹۴.۰۰	۹۵.۷۱	۹۲.۵۷
سرقت علمی با ابهام	۹۳.۷۳	۸۶.۲۶	۸۹.۸۴
کل پیکره	۹۰.۷۴	۸۹.۶۶	۹۰.۰۰

- نتایج بدست آمده برای تشخیص سرقت علمی اسناد انگلیسی
- ۱- بررسی تاثیر چند پارامتر مهم در تشخیص سرقت علمی نوع خلاصه
- ۲- همه سطوح ترازبندی متن

تاثیر پارامترهای مختلف در تشخیص سرقت علمی با ابهام خلاصه

- تاثیر جهت‌دار بودن گراف متن در تشخیص سرقت علمی نوع خلاصه اسناد انگلیسی

نوع گراف	دقت	فراخوانی	معیار F
گراف جهت‌دار	۸۷.۴۴	۸۰.۵۹	۸۳.۸۸
گراف غیر جهت‌دار	۷۸.۷۶	۸۱.۴۰	۸۰.۳۶

تاثیر پارامترهای مختلف در تشخیص سرقت علمی با ابهام خلاصه–۱

- تاثیر حداقل آستانه مشابهت گره‌ها در تشخیص سرقت علمی نوع خلاصه اسناد انگلیسی

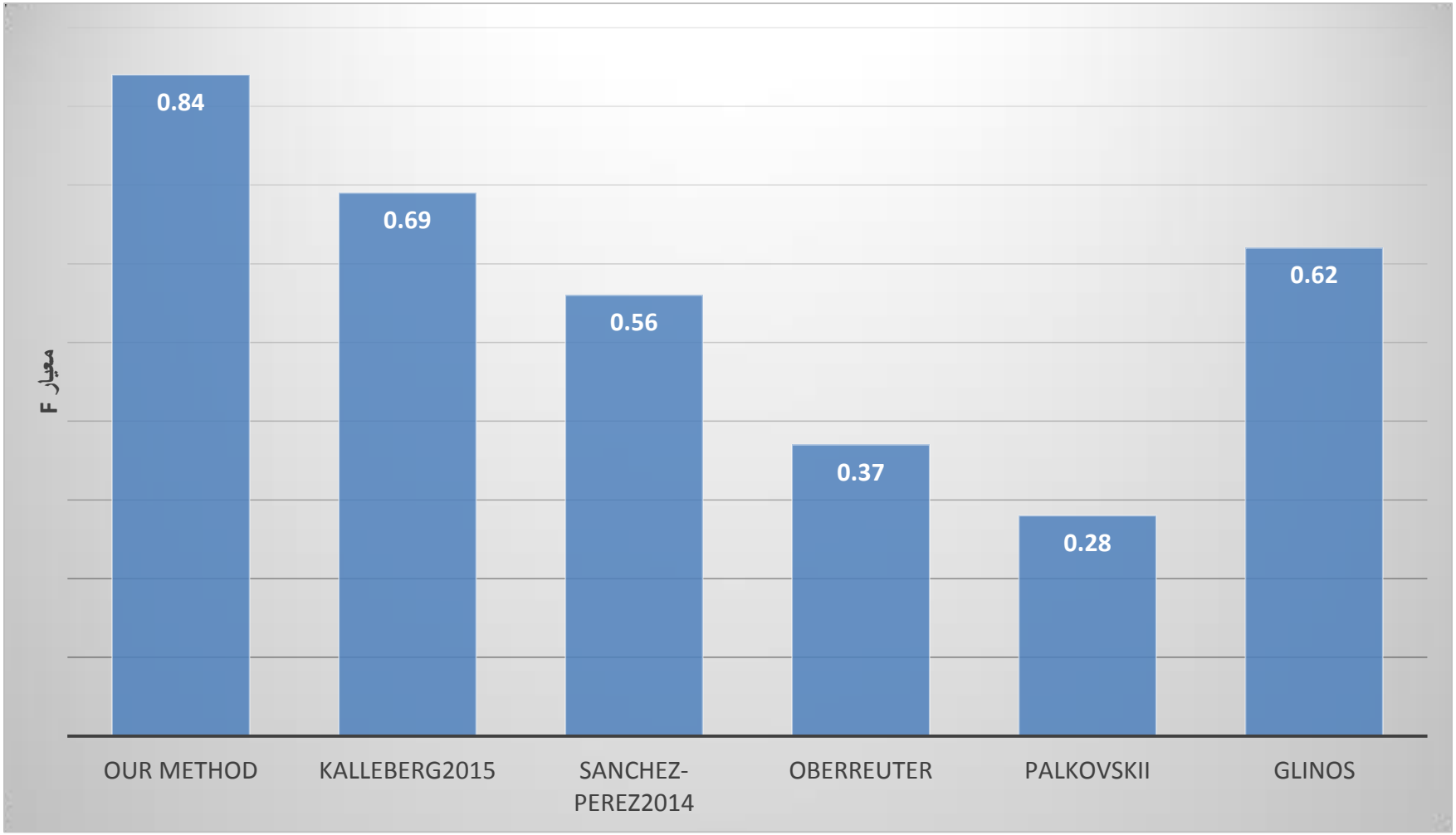
نوع گراف	دقت	فراخوانی	معیار F
α=0.4	۶۱.۷۷	۸۴.۴۴	۷۱.۳۴
α=0.6	۸۷.۴۴	۸۰.۵۹	۸۳.۸۸
α=0.8	۸۷.۷۵	۶۰.۹۳	۶۴.۴۵

تاثیر پارامترهای مختلف در تشخیص سرقت علمی با ابهام خلاصه–۲

- تاثیر نوع جداسازی متن درتشخیص سرقت علمی نوع خلاصه اسناد انگلیسی

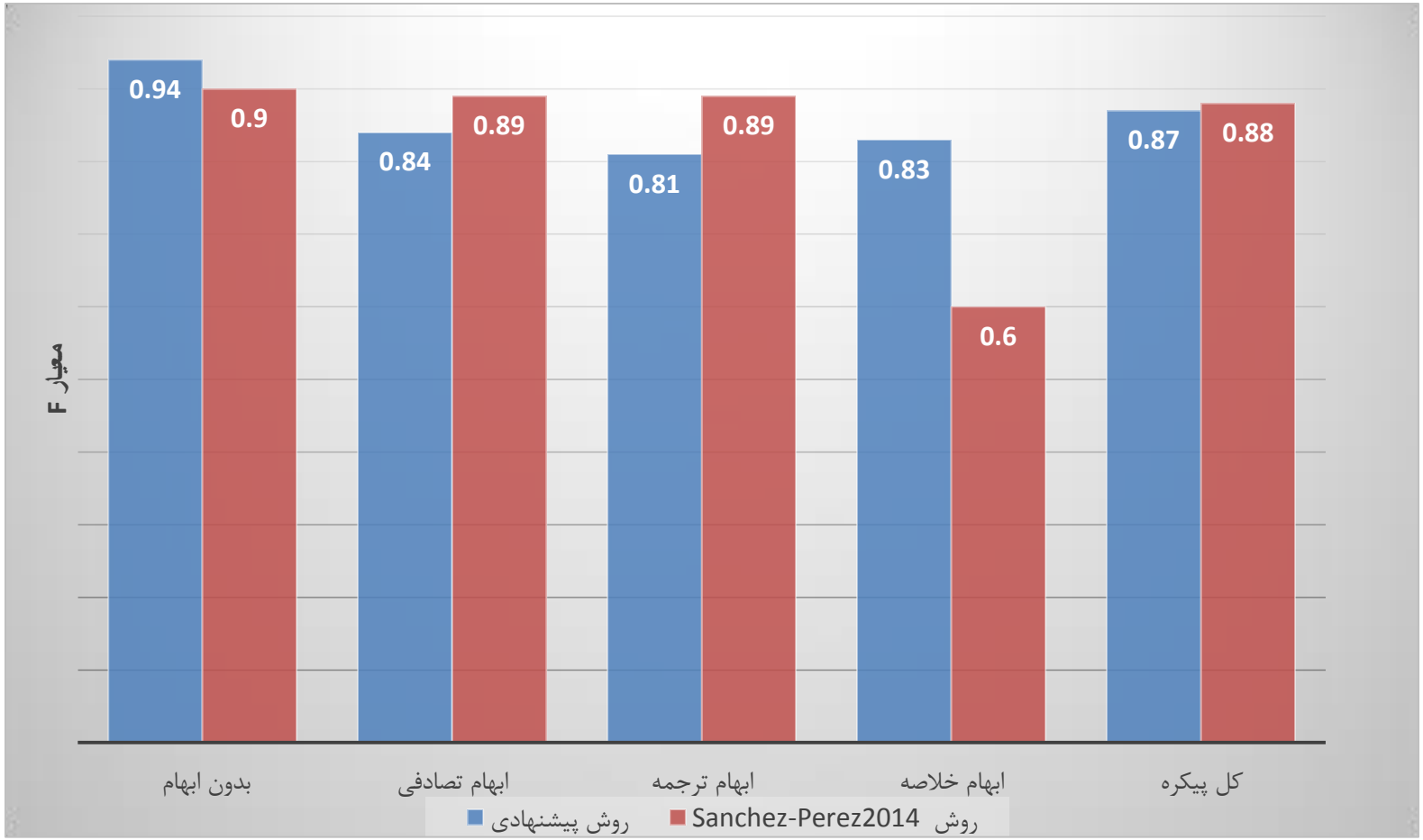
نوع گراف	دقت	فراخوانی	معیار F
تبدیل متن به مجموعه‌ای از پاراگراف‌ها	۸۷.۴۴	۸۰.۵۹	۸۳.۸۸
تبدیل متن به مجموعه‌ای از جملات	۹۱.۷۱	۱۵.۱۱	۲۷.۱۷
تبدیل متن به بخش‌هایی با طول ۱۰۰۰ کاراکتر	۶۶.۵۶	۱۴.۳۱	۲۳.۵۶
تبدیل متن به بخش‌هایی با طول ۵۰۰کاراکتر	۷۶.۹۴	۸.۸۱	۱۵.۸۲
تبدیل متن به بخش‌هایی با طول ۳۰۰ کاراکتر	۸۳.۳۷	۷.۵۰	۱۳.۹۰

مقایسه عملکرد روش پیشنهادی در تشخیص سرقت علمی با ابهام خلاصه با روش‌های پیشین



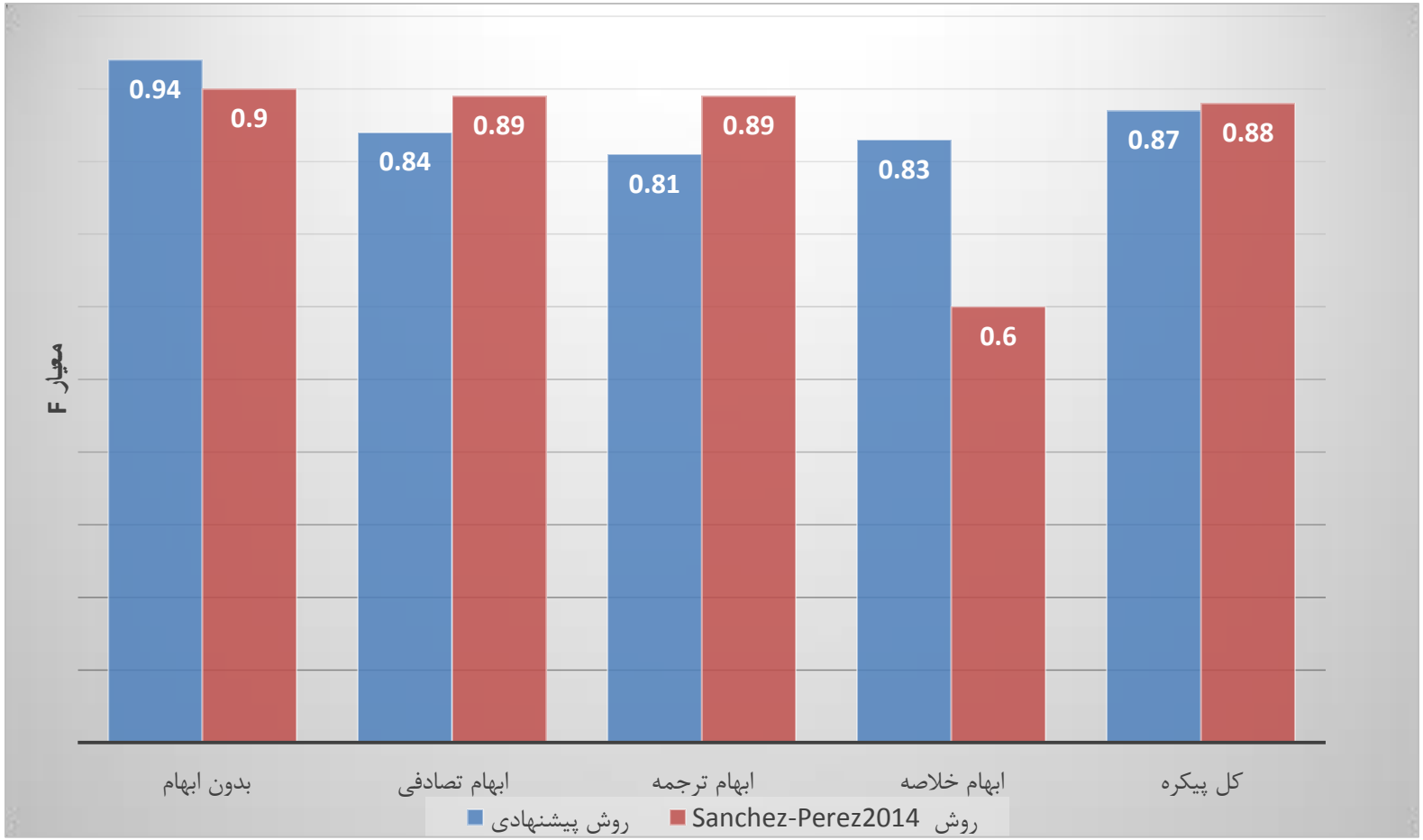
ارزیابی– نتایج بخش ترازبندی سند اسناد انگلیسی

- مجموعه داده آموزش



ارزیابی– نتایج بخش ترازبندی سند اسناد انگلیسی

- مجموعه داده آزمایش



نتایج اجرای روش برای مجموعه داده اسناد فارسی PAN2016

- مجموعه داده آموزش:

جدول ۲- نتایج به‌دست آمده از اجرای روش پیشنهادی روی مجموعه داده **PAN2016**

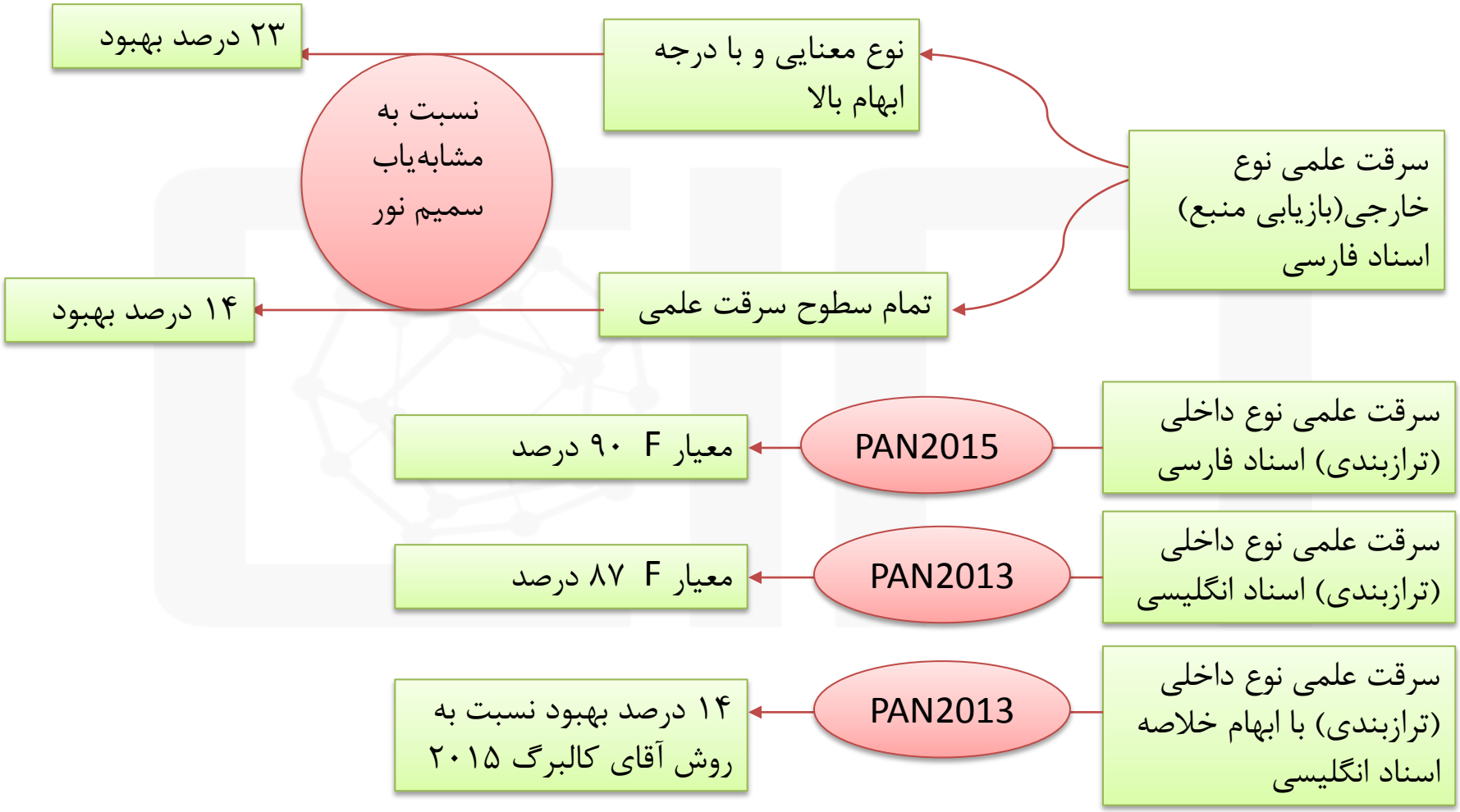
	دقت	فراخوانی	معیار F
نتایج	۹۰	۸۸.۵	۸۹.۳

نتایج اجرای روش برای مجموعه داده اسناد فارسیPAN2016

- مجموعه داده آزمایش:

Rank	Team	Plagdet	Granularity	Precision	Recall
1	Fatemeh Mashhadi, Mehrnoush Shamsfard Shahid Beheshti University, NLP Research Lab	**0.92204**	1.00146	0.92688	0.91919
2	Hadi Veisi, Kayvan Bijari, Kiarash Zahirnia, Erfaneh Gharavi University of Tehran, Data & Signal processing Lab	**0.90593**	1	0.95927	0.85820
3	Mozhgan Momtaz, Kayvan Bijari, Davood Heidarpour University of Tehran, COIN Lab	**0.87103**	1	0.89258	0.85049
4	Mahdi Niknam University of Qom	**0.83015**	1.03968	0.92034	0.79602
5	Faezeh Esteki, Faramarz Safi Esfahani Najafabad Branch, Islamic Azad University	**0.80083**	1.0	0.93337	0.70124

نتیجه‌گیری


سرقت علمی نوع خارجی(بازیابی منبع) اسناد فارسی
نوع معنایی و با درجه ابهام بالا
۲۳ درصد بهبود
تمام سطوح سرقت علمی
۱۴ درصد بهبود
نسبت به مشابه‌یاب سمیم نور
سرقت علمی نوع داخلی (ترازبندی) اسناد فارسی
PAN2015
معیار F ۹۰ درصد
سرقت علمی نوع داخلی (ترازبندی) اسناد انگلیسی
PAN2013
معیار F ۸۷ درصد
سرقت علمی نوع داخلی (ترازبندی) با ابهام خلاصه اسناد انگلیسی
PAN2013
۱۴ درصد بهبود نسبت به روش آقای کالبرگ ۲۰۱۵

کارهای آتی

- بهبود دقت سرقت علمی معنایی با نسخه توسعه یافته شبکه واژگان فارس‌نت
- انتخاب گره‌های موثرتر به کمک پیکره اسامی خاص
- بهبود دقت به کمک از بین بردن اثر کلمات مرکب پرتکرار زبان فارسی
- تعمیم روش به تشخیص سرقت علمی چند زبانه

مراجع – ۱

[6] Zini, Manuel, Fabbri, Marco, Moneglia, Massimo, and Panunzi, Alessandro, Plagiarism detection through multilevel text comparison, Second International Conference, IEEE, 2006.

[7] Potthast, Martin, Gollub, Tim, Hagen, Matthias, Tippmann, Martin, Kiesel, Johannes, Rosso, Paolo, Stamatatos, Efstathios, and Stein, Benno. Overview of the 5th International Competition on Plagiarism Detection. in Forner, Pamela, Navigli, Roberto, and Tufis, Dan, eds. , Working Notes Papers of the CLEF 2013 Evaluation Labs, September2013.

[8] Zager, Laura. Graph similarity and matching. Ph.D. thesis, Massachusetts Institute of Technology, 2005.

[9] Blanco, Roi and Lioma, Christina. Graph-based term weighting for information retrieval. Information retrieval, 15(1):54–92, 2012.

[10] Osman, Ahmed Hamza, Salim, Naomie, and Binwahlan, Mohammed Salem. Plagiarism detection using graph-based representation. arXiv preprint arXiv:1004.4449, 2010.

[11] K. e. a. Khoshnavataher, "Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation," PAN-CELF, 2015.

مراجع–۲

[1] N.Kumar, A graph based automatic plagiarism detection technique to handle artificialword reordering and paraphrasing, Computational Linguistics and Intelligent Text Processing,p. 481–494, 2014.

[2] Sonawane, S. S., & Kulkarni, P. A, "Graph based Representation and Analysis of Text Document: A Survey of Techniques.," in International Journal of Computer Applications, 96(19), 2014.

[3] Kalleberg, Rune Borge. Towards detecting textual plagiarism using machine learning methods. 2015.

[4] Sanchez-Perez, Miguel A, Sidorov, Grigori, and Gelbukh, Alexander F. A winning approach to text alignment for text reuse detection at pan 2014. in CLEF (Working Notes), pp. 1004–1011, 2014.

[5] Oberreuter, Gabriel, and Juan D. VeláSquez, Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style, Expert Systems with Applications 40.9: 3756-3763, 2013.

پایان

از توجه شما عزیزان سپاسگزارم

سرانجام روزی به حکمت
همه اتفاقات زندگیتان
پی خواهید برد. پس
فعلاً به سردرگمی ها
بخندید، از میان اشکها
لبخند بزنید، و همواره
به خودتان یادآوری
کنید که پشت هر
حادثه ای دلیلی نهفته
است.