Introducing a Novel Method for Automatic Facet Extraction in Faceted Search (case study: Gynaecology and Obstetrics Domain)

Abdolhossein Farajpahlou
PhD in Knowledge & Information Science; Professor; School of Education & Psychology; Shahid Chamran University of Ahvaz; Ahvaz, Iran Email: farajpahlou@gmail.com

Farideh Osareh
PhD in Knowledge & Information Science; Professor; School of Education & Psychology; Shahid Chamran University of Ahvaz; Ahvaz, Iran Email: osareh.f@scu.ac.ir

Seyed Mostafa Fakhrahmad
PhD in Computer Engineering; Associate Professor; Department of Computer Science and Engineering & IT; Shiraz University; Shiraz, Iran Email: mfakhrahmad@yahoo.com

Leila Dehghani*
PhD in Knowledge & Information Science; Assistant Professor; Medical Librarianship Group; Department of Paramedical Medicine; Bushehr University of Medical Sciences; Bushehr, Iran; Email: leiladehghani@yahoo.com

Iranian Journal of Information Processing and Management

Received: 10, Apr. 2021 **Accepted: 28, Aug. 2021**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 37 | No. 3 | pp. 807-838
Spring 2022
https://doi.org/10.35050/JIPM010.2022.788

Abstract: In this research a new algorithm for facet extraction has been developed and introduced, which provides the experimental possibility of identifying facets based on a literary warrant. In the field of automatic facet extraction two main ideas were considered by reviewing the researches. The first idea is that the facet appears in the context. Therefore, to identify the facet in a corpus, its context must be examined. The second idea is that the facet is the focal point in a lexical tree that is neither very general nor very specific.

Based on these two ideas, first, the corpus in the medicine area and the obstetrics and gynaecology domain was prepared. The research team selected three corpora from the literary warrant and used the abstract and title of the collection of articles in top 20 journals of the field to create a contextual corpus. This collection contained 167071 documents. 2000 articles were randomly selected to create the origin corpus. The third body is the lexical corpus. The proper words of the corpus were extracted using a web-based service. The output contained 514 words. Duplicate words were removed and finally, 480 important words were identified.

* Corresponding Author

Then, the words were expanded in the contextual corpus with the help of the supervisor (Mesh) and then-candidate dissertations were extracted based on the two conditions of frequency-based Shifting and rank-based Shifting. Finally, using the three rules of specificity, substitution, and generality, the identified facets were modified and named. Finally, 26 facets were identified in the domain of gynaecology and obstetrics. Comparing the proposed algorithm with other algorithms, it was found that the combination of statistical approach and tree pruning can have better results than purely statistical approach or tree pruning. Also, the comparison of the output facets of the algorithm with the traditional facets in this obstetrics and gynaecology domain showed that the output of the algorithm is smaller and more useful for browsing information retrieval tools. Also, in this study was specified that specialized domain facets are different from general facets and can be redefined independently, but the results cannot be generalized to all medical domains and other researches are needed to be done in other fields.

Keywords: Data Retrieval, Facet, Faceted Search, Automatic Facet Extraction

ارائۀ روشی نو برای استخراج خودکار چهریزه‌ها در جست‌وجوهای چهریزه‌ای (مورد مطالعه: حوزۀ زنان و زایمان)

عبدالحسین فرج‌پهلو
دکتری علم اطلاعات و دانش‌شناسی؛ استاد؛ دانشگاه شهید چمران اهواز؛ اهواز، ایران؛
farajpahlou@gmail.com

فریده عصاره
دکتری علم اطلاعات و دانش‌شناسی؛ استاد؛ دانشگاه شهید چمران اهواز؛ اهواز، ایران؛
osareh.f@ scu.ac.ir

سید مصطفی فخراحمد
دکتری مهندسی رایانه؛ دانشیار؛ دانشگاه شیراز؛ شیراز، ایران mfakhrahmad@yahoo.com

لیلا دهقانی
دکتری علم اطلاعات و دانش‌شناسی؛ استادیار؛ دانشگاه علوم پزشکی بوشهر؛ بوشهر، ایران؛
پدیدآور رابط leiladehghani@yahoo.com

پژوهش‌نامۀ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۰/۰۱/۲۱ **پذیرش:** ۱۴۰۰/۰۶/۰۶ مقاله برای اصلاح به مدت ۱۵ روز نزد پدیدآوران بوده است.

چکیده: هدف این پژوهش ابداع و معرفی الگوریتمی نو برای استخراج چهریزه‌هاست که امکان شناسایی تجربی چهریزه‌ها را با کمک پشتوانۀ انتشاراتی فراهم می‌کند. الگوریتم پیشنهادی بر مبنای دو ایده شکل گرفته است: ایدۀ اول اینکه چهریزه در بافت بروز پیدا می‌کند. بنابراین، برای تشخیص چهریزه در یک بدنۀ متنی بایستی بافت یا بستر آن مورد بررسی قرار گیرد و ایدۀ دوم این است که چهریزه نقطۀ تمرکز در یک درخت واژگانی است که نه بسیار عام و نه بسیار خاص است.
در حوزۀ پزشکی، دامنۀ زنان و زایمان به‌عنوان بستر آزمون انتخاب گردید. سه پیکرۀ متنی از درون پشتوانۀ انتشاراتی انتخاب شد. پیکرۀ بستر، از چکیده و عنوان مجموعه مقالات موجود در ۲۰ مجلۀ برتر حوزه انتخاب شد که دربرگیرندۀ ۱۶۷۰۷۱ سند بود. پیکرۀ دوم، پیکرۀ منشأ بود که ۲۰۰۰ مقاله به‌صورت تصادفی از پیکرۀ بستر انتخاب شد. پیکرۀ سوم، پیکرۀ واژگانی است که با استفاده از یک سرویس تحت وب و معیار رتبه‌بندی واژگان LIDF-value استخراج گردید. خروجی حاصل دربرگیرندۀ ۵۱۴ واژه بود. واژگان تکراری حذف شدند و سرانجام، ۴۸۰ واژۀ مهم شناسایی شد.

نشریۀ علمی | رتبۀ 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
دورۀ ۳۷ | شمارۀ ۳ | ص‌ص ۸۰۷-۸۳۸
بهار ۱۴۰۱
https://doi.org/10.35050/JIPM010.2022.788

سپس، واژگان در پیکرهٔ بستر با کمک مجموعهٔ راهنما یعنی «مش» بسط داده شد و پس از آن، بر اساس دو شرط انتقال مبتنی بر تکرار یعنی بیشتر بودن اسناد مرتبط با واژه در بستر نسبت به منشأ و انتقال مبتنی بر رتبه یعنی رشد رتبهٔ موجود واژه در پیکرهٔ بستر نسبت به منشأ که نشان‌دهندهٔ عام شدن واژه است، چهریزه‌های کاندید استخراج شدند. سرانجام، با استفاده از سه قاعدهٔ اخص بودن، جایگزنی و اعم بودن، چهریزه‌های شناسایی‌شده اصلاح و نام‌گذاری شدند. در نهایت، ۲۶ چهریزه به‌عنوان چهریزه‌های حوزهٔ زنان و زایمان شناسایی شدند.
با مقایسهٔ الگوریتم پیشنهادی با دیگر الگوریتم‌ها مشخص شد که ایجاد سه افراز (افراز منشأ و بدنهٔ متنی و افراز برای شناسایی واژگان مهم) و مقایسهٔ رفتار واژه در آن‌ها و سپس، ایجاد درخت بر اساس چهریزه‌های کاندید، یعنی ترکیب رویکرد آماری و هرس درخت می‌تواند نتایج مناسب‌تری نسبت به رویکرد صرفاً آماری یا هرس درخت داشته است. همچنین، مقایسهٔ چهریزه‌های خروجی از الگوریتم و چهریزه‌های سنتی در این زمینه نشان داد که چهریزه‌های خروجی الگوریتم، خردتر و برای مرور در ابزارهای بازیابی اطلاعات مفیدتر هستند. همچنین، در این پژوهش مشخص شد که چهریزه‌های دامنهٔ تخصصی از چهریزه‌های عمومی در حوزهٔ پزشکی متفاوت است و مستقل از آن‌ها قابل شناسایی و تعریف است، اما نمی‌توان نتایج را به تمامی دامنه‌های پزشکی تعمیم داد و نیاز است که پژوهش‌هایی در دیگر حوزه‌ها صورت گیرد.

کلیدواژه‌ها: بازیابی اطلاعات، چهریزه، جست‌وجوی چهریزه‌ای، استخراج خودکار چهریزه

۱. مقدمه

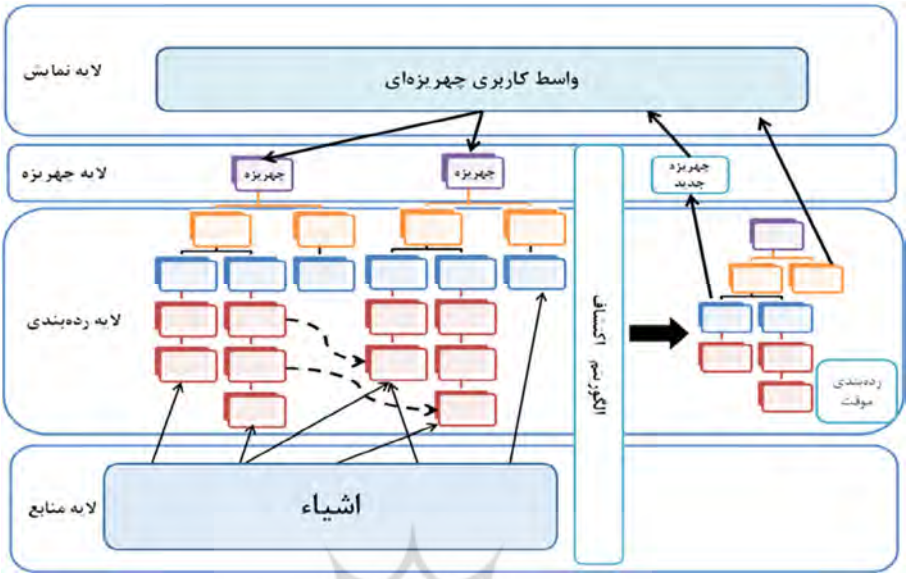
مفهوم چهریزه در حوزهٔ کتابداری و اطلاع‌رسانی قدمتی طولانی دارد. رده‌بندی دهدهی جهانی[1] (۱۹۰۷)، رده‌بندی کولُن[2] (۱۹۳۳)، رده‌بندی کتابشناختی بلیس[3] (۱۹۴۰) از نمونه‌های برجستهٔ به‌کارگیری چهریزه هستند. رویکرد تحلیل چهریزه‌ای از حدود اوایل قرن بیستم تا سال ۱۹۹۰ میلادی بر مبنای نظام منطقی (پیشینی) طبقه‌بندیِ علوم پیش رفته است. اما از آن سال به بعد، به‌دلیل گسترش توانایی‌های رایانه‌ای، دیدگاه منطقی جای خود را به دیدگاه محاسباتی و کاربرمدار (پسینی) سپرد (Farajpahlou et al. 2019). در دیدگاه کاربرمدار، تعیین چهریزه وابسته به نیاز کاربر است و برای دستیابی به چهریزه‌ها در حوزهٔ موضوعی بایستی از پشتوانهٔ کاربری استفاده نمود. هدف اصلی این رویکرد، شناسایی نیازهای اطلاعاتی کاربران و استخراج چهریزه‌ها از آن (به‌طور مستقیم با کمک کاربران) برای طراحی نظام بازیابی اطلاعات است. افزون بر این، استفاده از چهریزه‌های

1. Universal Decimal Classification (UDC)
2. Colon Calcification
3. Bliss Bibliographic Classification (BC)

استخراج‌شده به توسعه و یکپارچگی هستی‌شناسی‌های تخصصی نیز کمک می‌کند (Farajpahlou et al. 2020).

رویکرد دیگر، رویکرد محاسباتی است. در حال حاضر، این رویکرد بر نقش راهبری چهریزه در سیستم‌های بازیابی اطلاعات و «جست‌وجوی چهریزه‌ای» تمرکز یافته است (Hudon 2020). جست‌وجوی چهریزه‌ای جست‌وجویی است که بر اساس چهریزه‌های تعریف‌شده این امکان را به کاربر می‌دهد که به‌صورت مرحله‌ای چهریزه‌های مختلف را در جست‌وجو وارد نموده و نتایج را پالایش نماید. در این روش، چهریزه راهی برای طبقه‌بندی اطلاعات است که جست‌وجوگران را برای پالایش جست‌وجو از طریق گروه‌بندی اسناد یاری می‌رساند. همچنین، در جست‌وجوی کلیدواژه‌ای نتیجهٔ صفر وجود دارد، اما در جست‌وجوی چهریزه‌ای، جست‌وجو در بافت[1] انجام می‌شود و بنابراین، نتیجهٔ صفر نخواهد داشت (Zheng, Zhang and Feng 2013). از سوی دیگر، جست‌وجوی چهریزه‌ای سبب ایجاد پرس‌وجوهای پیچیده‌تری می‌شود. افزون بر این، چهریزه‌ها بسیار انعطاف‌پذیر هستند، زیرا آن‌ها تنها نقاطی را که پژوهشگر می‌خواهد جست‌وجو می‌کنند (Sacco and Tzitzikas 2009). سیستم جست‌وجوی چهریزه‌ای دربرگیرندهٔ انواع رابط‌های چهریزه‌ای، مکانیزم‌های استخراج چهریزه، رده‌بندی چهریزه‌ای و روابط افقی و عمودی آن و پایگاه اطلاعاتی چهریزه‌ای است. شکل ۱، نشان‌دهندهٔ اجزای سیستم جست‌وجوی چهریزه‌ای است. در این شکل، لایهٔ منابع از پشتوانهٔ انتشاراتی استخراج می‌شود. در لایهٔ رده‌بندی دو گونه رده‌بندی چهریزه‌ای ایجاد می‌شود: رده‌بندی چهریزه‌ای حاصل از رویکرد منطقی که شاخص‌ترین آن رده‌بندی «کولُن» است، و رده‌بندی موقت حاصل از الگوریتم‌های رایانه‌ای که برای نمونه می‌توان به رده‌بندی پویای Sacco (2000) و پروژهٔ «فلامنکو» (Hearst 2006) اشاره کرد. در لایهٔ بالاتر از رده‌بندی، لایهٔ چهریزه قرار دارد که در آن، چهریزه‌ها به‌صورت سنتی یا به‌صورت الگوریتمی شناسایی می‌شوند و در بالاترین لایه، لایهٔ نمایش یا واسط کاربری چهریزه‌ای است که امروزه، در اکثر وب‌سایت‌های معتبر استفاده می‌شود. تمرکز این مقاله بر لایهٔ چهریزه و مکانیزم استخراج چهریزه است.

1. context



شکل ۱. سیستم جست‌وجوی چهریزه‌ای (پژوهش حاضر)

«یورلند» معتقد است که در مطالعات علوم کتابداری و اطلاع‌رسانی، مکانیزم‌های استخراج چهریزه‌ها فاقد منابع تجربی لازم هستند؛ چرا که در آن‌ها از یک نوع معرفت‌شناسی عقل‌گرا (سنتی)[1] برای استخراج چهریزه‌ها استفاده می‌شود. این نوع معرفت‌شناسی بر اساس طبقه‌بندی عمومی و منطقی شکل می‌گیرد. در حقیقت یک طبقه‌بندی غیرقابل تغییر و درونی را شکل می‌دهد که مفاهیم و ارتباط آن‌ها پیشینی است. چهریزه‌ها بر اساس منطق ارسطو (پیشینی) شناسایی و سعی می‌شود تمامی مفاهیم به‌گونه‌ای در قالب این چهریزه‌ها یا ابعاد جای گیرند. اما در محیط واقعی، مفاهیم و ارتباط آن‌ها از بالا (پیشینی) شکل نمی‌گیرد، بلکه بر اساس مشاهدات تجربی از پایین (استقرایی و پسینی) شکل می‌گیرد. از نگاه «یورلند» گرایش به سمت الگوریتم‌های استخراج چهریزه، حرکت به‌سوی تجربه‌گرایی است (Hjørland 2003). گرچه پژوهش‌های سال‌های اخیر در حوزهٔ استخراج خودکار چهریزه‌ها صورت گرفته است، ولی آن‌ها همچنان از کاربردهای عملی فاصله دارند. همچنین، به‌دلیل نگاه عقل‌گرای حاکم، چهریزه‌های شناسایی‌شده عمدتاً ثابت هستند و مستقل از دامنهٔ تخصصی تعریف می‌شوند. از سوی دیگر، «یورلند»

1. logical

معتقد است که چهریزه‌ها در یک دامنهٔ خاص تعریف نشده‌اند و پیشنهاد داده است که شناسایی چهریزه‌ها در دامنهٔ تخصصی آزمون شود (Hjørland 2013). این در حالی است که برخی دیگر اعتقاد دارند چهریزه‌ها مستقل از دامنهٔ تخصصی هستند (Zheng, Zhang and Feng 2013). از همین رو، در این مقاله به ابداع و معرفی الگوریتمی نو برای استخراج چهریزه‌ها و روابط آن‌ها پرداخته شده است که امکان تجربی برای شناسایی چهریزه‌ها با کمک پشتوانهٔ انتشاراتی را فراهم می‌کند. از سوی دیگر، با به‌کارگیری این الگوریتم در حوزهٔ موضوعی زنان و زایمان در دامنهٔ پزشکی به این پرسش پاسخ داده می‌شود که آیا چهریزه‌ها در حوزهٔ تخصصی از دامنهٔ موضوعی خود مستقل هستند؟

۲. پیشینهٔ پژوهش

الگوریتم‌های پیشنهادی برای استخراج چهریزه‌ها از مفاهیم نظریهٔ مجموعه‌ها، نظریهٔ گراف‌ها و نظریهٔ آماری زبانی استفاده کرده‌اند. تجارب حاصل از این الگوریتم‌ها در طراحی وب‌سایت‌های چهریزه‌ای، نرم‌افزارهای نمایه‌سازی چهریزه‌ای کتابخانه‌ها، توصیف هستی‌شناسی‌ها و رده‌بندی‌های چهریزه‌ای استفاده می‌شود. با مرور یک دورهٔ سی‌ساله از ۱۹۹۰ تا ۲۰۲۰ مجموعه‌ای از الگوریتم‌ها شناسایی شد که در ادامه، مطرح‌ترین آن‌ها در قالب جدول ۱، تشریح شده‌اند.

جدول ۱. نمونه‌های مطرح الگوریتم‌های استخراج چهریزه

زمان	نام الگوریتم	شرح	نقاط قوت	نقاط ضعف
(Hearst 2006; Stoica, Hearst and Richardson 2007)	الگوریتم Castanet	ایدهٔ اصلی مدل این است که در ازای تهیهٔ یک رده‌بندی بسیار بزرگ، رده‌بندی‌های کوچک مبتنی‌بر چهریزه تهیه و بین آن‌ها روابط را تعریف کنیم. در این صورت، جست‌وجو می‌تواند هدایت‌کنندهٔ کاربر اصلی باشد. پس از مدل اولیه در سال ۲۰۰۷ اصلاحاتی صورت گرفت و الگوریتم پیشنهادی ارائه گردید که castanet نام دارد. ورودی این الگوریتم یک بدنهٔ متنی است. در گام اول، واژگان مهم از بدنهٔ متنی استخراج می‌شود. ملاک انتخاب، توزیع واژگان در بدنهٔ متنی است. واژه‌هایی که بیش از آستانه در نظر گرفته شده بودند، به‌عنوان واژهٔ هدف شناسایی می‌شوند. در گام دوم، درخت اصلی ساخته می‌شود. ابتدا، واژگان انتخاب شده، ابهام‌زدایی می‌شوند و دامنه‌ای که به آن تعلق دارد در شبکهٔ واژگان wordnet شناسایی می‌شود و در نهایت، تمامی مسیرها برای واژگان در شبکهٔ واژگان wordnet شناسایی شده و با یکدیگر ادغام می‌گردند و سبب ایجاد یک درخت واحد می‌شوند. در گام سوم، درخت اصلی تقویت می‌شود و در گام آخر با بررسی مسیرهای درخت، برخی از واژه‌ها در نقطهٔ تمرکز قرار می‌گیرند که چهریزه را شکل می‌دهند.	ویژگی این الگوریتم استفاده از الگوریتم هرس درخت است که نسبت به گراف‌ها پیاده‌سازی آسان‌تری دارد.	ایراد اساسی این الگوریتم عدم توجه به بافت است. این الگوریتم نمی‌تواند بافت را تشخیص دهد و به‌شدت وابسته به شبکهٔ واژگان wordnet است.

زمان	نام الگوریتم	شرح	نقاط قوت	نقاط ضعف
(Dakka 2008; Dakka, Ipeirotis and Sacco 2009)	الگوریتم مدل داکا	این الگوریتم، ارائهٔ راهکاری برای تبدیل رده‌بندی‌های غیرچهریزه‌ای و یا چهریزه‌ای قدیمی به یک رده‌بندی چهریزه‌ای مبتنی بر رویکرد محاسباتی است. در متن‌های غیرساخت‌یافته تعیین چهریزه‌ها نیازمند یک گام پیش‌پردازش است. متن از طریق ابزارهای پردازش زبان طبیعی به مجموعه‌ای از واژگان تبدیل می‌شود. بنابراین، نیازمند استفاده از منابع بیرونی مانند wordnet، google directory، Wikipedia و رده‌بندی‌های عمومی و تخصصی است. بنابراین، مفهوم واژه‌های استخراج‌شده بر اساس منابع بیرونی توسعه پیدا می‌کند یا «بافت آگاه» می‌شود. سپس، با استفاده از تحلیل تکرار واژه‌ها فهرستی مرتب از واژگان استخراج می‌شود و با تعیین تابع درست‌نمایی و تعیین آستانهٔ مناسب فهرست نهایی چهریزه‌ها استخراج می‌گردد. در این پژوهش روش تعیین چهریزه با استفاده از عامل انسانی نیز صورت گرفته و خروجی دو روش با هم مقایسه شده است که نشان‌دهندهٔ نزدیکی نتایج است. اما بایستی به این نکته اشاره کرد که این مدل در دامنهٔ تخصصی صورت نگرفته و متن‌ها از آرشیو روزنامهٔ «نیویورک‌تایمز» به‌دست آمده است.	ویژگی مهم، «بافت آگاهی» این روش است. همچنین، این روش نیاز به مربی ندارد.	نبودِ درخت هرس برای اصلاح چهریزه‌های استخراج‌شده است.

زمان	نام الگوریتم	شرح	نقاط قوت	نقاط ضعف
(Li et al. 2010)	الگوریتم Facetedpedia	این الگوریتم بر اساس نظریهٔ گراف‌ها بنا شده و به‌صورت یک مجموعهٔ جبری از روابط تعریف شده است. هدف آن کشف پویای چهریزه‌ها برای رابط کاربری بر اساس پرس‌وجوست. ورودی این الگوریتم مجموعه‌ای از مقالات مرتب‌شدهٔ «ویکی‌پدیا»ست که از طریق جست‌وجو بر اساس لغت کلیدی ایجاد شده است. این جست‌وجو به ایجاد یک گراف ریشه‌دار می‌انجامد که گره‌های گراف، یک موضوع اطلاعاتی در «ویکی‌پدیا» هستند. هر مقاله به مجموعه‌ای از مقالات دیگر پیوند دارد و به آن مقالات «دارای ویژگی»[1] می‌گویند. هر مقالهٔ «دارای ویژگی» می‌تواند به یک یا چند گره (موضوع اطلاعاتی) متصل شود. چهریزه، مجموعه‌های ممکن از سلسله‌مراتب است که زیرمجموعهٔ سلسله‌مراتب مادر (ورودی الگوریتم) است. مسیر هدایت به‌گونه‌ای تعریف می‌شود که مقالهٔ هدف از مسیر طبقه‌بندی و مقالات دارای ویژگی قابل دستیابی باشد. چون مقالات دارای ویژگی به یک یا چند گره تخصیص داده شده‌اند، مسیرهای متفاوتی برای دستیابی به مقالهٔ هدف وجود دارد.	استفاده از نظریهٔ گراف‌هاست که پویایی نتایج را افزایش می‌دهد.	ایراد اساسی این الگوریتم استفاده از سند به‌عنوان ورودی است و از استخراج واژگان استفاده نمی‌کند و مورد دوم، هیچ مکانیزمی برای پالایش خروجی یعنی چهریزه‌ها ندارد و وابسته به ساختار سلسله‌مراتبی و ارتباطی بین اسناد «ویکی‌پدیا»ست.
(Basu Roy 2011)	تشخیص چهریزه Dynacet	این الگوریتم با هدف ایجاد یک رابط کاربری چهریزه‌ای برای پایگاه‌های دادهٔ بسیار بزرگ تهیه شده است. این نرم‌افزار دارای یک بخش تولید چهریزه‌ها و یک بخش ساخت درخت چهریزه‌ای است. در این الگوریتم برای یک پایگاه اطلاعاتی با n رکورد و m صفت، یک درخت تصمیم ایجاد می‌شود. این درخت تصمیم نشان‌دهندهٔ انواع پرس‌وجوهای ممکن در پایگاه است. بدین معنا که برای هر صفت تمامی حالات پرس‌وجو (ساده و ترکیبی) شناسایی و با استفاده از یک تابع ابهام‌زدایی بهترین حالت پرس‌وجو، یعنی کمینه‌ترین راه برای دستیابی به سند تشخیص داده می‌شود. بنابراین، صفتی که سبب جست‌وجو با عملکرد بالا می‌شود، به‌عنوان چهریزه معرفی می‌شود.	کاربر این الگوریتم در پایگاه‌های دادهٔ رابطه‌ای	ایراد اساسی این الگوریتم استفاده از یک پایگاه اطلاعاتی ساخت‌یافته است؛ یعنی از استخراج واژگان استفاده نمی‌کند، و مورد دوم هیچ مکانیزمی برای پالایش خروجی یعنی چهریزه‌ها ندارد.

زمان	نام الگوریتم	شرح	نقاط قوت	نقاط ضعف
(Komamizu 2015)	الگوریتم مبتنی‌بر گراف داده‌ها	استخراج چهریزه بر اساس الگوهای تکراری در گراف داده‌ها، عصارهٔ اصلی الگوریتم پیشنهادی است. گراف دادهٔ جامع ورودی الگوریتم است. گراف داده با استفاده از روشی به زیرگراف‌هایی تبدیل می‌شود که هر زیرگراف دربرگیرندهٔ یک کلاس است. صفات کلاس رأس‌های گراف (در اینجا مفاهیم)، و روش یا متد کلاس یال‌های گراف (روابط بین مفاهیم) هستند. معیار مفید بودن چهریزه به تکرار واژه در صفات کلاس‌های گوناگون و میزان ارتباط بین صفات (رأس گراف) بازمی‌گردد. برای نمونه، اگر در مجموع، ۴۰ زیرگراف از یک شبکهٔ استنادی (گراف اصلی) استخراج شود و نام نویسنده سند (صفت) بیش از ۲۰ بار (حد آستانه) تکرار شده باشد، نویسنده به‌عنوان چهریزه انتخاب می‌گردد.	–	باید خاطرنشان کرد که این روش برای متون غیرساخت‌یافته امتحان نشده است و کاربرد آن بیشتر در پایگاه‌های داده ساخت‌یافته و یا فرمت‌های نیمه‌ساخت‌یافته مانند XML است.
(Siddiqui et al. 2016)	FacetGist	هدف الگوریتم این است که به‌طور خودکار هر سند را با مجموعه‌ای از مفاهیم که چهریزه‌های اصلی هستند، برچسب‌گذاری کند. چالش عمده در این الگوریتم، استخراج مفهوم، تطبیق مفهوم با چهریزه، و ابهام‌زدایی از چهریزه است. برای مقابله با این چالش‌ها چارچوب FacetGist ارائه شد. استخراج چهریزه شامل ساخت یک شبکهٔ ناهمگن مبتنی بر گراف مفاهیم است. سپس، یک مسئلهٔ بهینه‌سازی ترسیم شده و یک الگوریتم کارآمد برای انتشار برچسب مبتنی بر گراف تهیه شد. یال‌هایی که بیشترین رابطه را دارند، برای تخمین چهریزه‌ها استفاده شدند. نتایج تجربی در اسناد فنی شرکت «مایکروسافت» نشان داد که استخراج چهریزه می‌تواند به بهبود بیش از ۲۵ درصد در دقت و فراخوانی اسناد منجر شود.	استفاده از شبکهٔ ناهمگن و وجود بررسی تجربی در یک محیط واقعی	چهریزه‌ها را معرفی نمی‌کند، بلکه تنها به نتایج خروجی بازیابی اطلاعات با استفاده از چهریزه‌ها تأکید دارد.

زمان	نام الگوریتم	شرح	نقاط قوت	نقاط ضعف
(Mauro et al. 2020)	الگوریتم مبتنی‌بر هستی‌شناسی	در این پژوهش، از نمایش معنایی هستی‌شناسی در دامنهٔ جغرافیایی محدود باستانی در «ایتالیا» استفاده شد. این هستی‌شناسی برای بازیابی اطلاعات (به‌عنوان پشتیبان) از منابع داده‌های خارجی استخراج شد. تعداد چهریزه‌ها مرتبط با یک گروه دادهٔ خاص به غنای فراداده آن بستگی دارد و می‌تواند نسبتاً زیاد باشد. با این حال، همهٔ این برچسب‌ها به یک اندازه مفید نیستند؛ بعضی از آن‌ها به‌ندرت قابل استفاده هستند. موارد دیگر معرف شناسه‌ها هستند و بنابراین، نمی‌توانند مورد را پشتیبانی کنند. بنابراین، از یک تابع هزینه استفاده شد. تابع هزینه، اجرای یک پرس‌وجو با استفاده از هر برچسب هستی‌شناسی را محاسبه می‌کند و سپس، با مرتب‌کردن آن‌ها برچسب‌هایی را که کمترین هزینه را داشته‌اند، به‌عنوان چهریزه انتخاب می‌کند. تابع هزینه از قانون آنتروپی استفاده می‌کند.	به‌کارگیری هستی‌شناسی و استفاده از روابط معنایی	از بدنهٔ متنی استفاده نمی‌کند.

۳. روش پژوهش

این پژوهش از نوع کاربردی و روش آن، مدل‌سازی آماری و هرس‌گراف است. با بررسی پژوهش‌های صورت‌گرفته در زمینهٔ استخراج اتوماتیک چهریزه، دو ایدهٔ اصلی مد نظر قرار گرفت: ایدهٔ اول مبتنی بر این عقیده است که چهریزه در بافت بروز پیدا می‌کند. بنابراین، برای تشخیص چهریزه در یک بدنهٔ متنی[1] بایستی بافت یا بستر آن مورد بررسی قرار گیرد؛ بدین معنا که واژگان مهم در بدنهٔ متنی شناسایی شده و سپس، واژگان مرتبط با آن‌ها در بافت شناسایی شوند. مقایسهٔ این دو مجموعه (مجموعهٔ واژگان مهم بدنهٔ متنی و بستر) سبب ایجاد چهریزه‌ها می‌گردد. مدل‌های «داکا»[2]، «فاسیت‌پیدیا»[3] و «فاست»[4] از این ایده تبعیت نموده‌اند. ایدهٔ دوم مبتنی بر این عقیده است که چهریزه نقطهٔ تمرکز در یک درخت واژگانی است که نه بسیار عام و نه بسیار خاص است؛ بدین معنا که میزان برگ‌ها و تعداد شاخه‌های سطح بالا و پایین آن نباید از یک آستانه بیشتر و از آستانهٔ دیگر کمتر باشد. میزان این آستانه بر اساس توزیع آماری واژگان و یا به‌صورت سعی و خطا[5] حاصل می‌شود. مدل‌های «اچ‌اف‌سی»[6] و «کاستانت»[7] نمونه‌هایی از این ایده هستند.

1. corpus
2. Daka
3. Facetedpedia
4. FaSet
5. trial and error
6. HFC
7. castanet

بر اساس این دو ایدهٔ تجربی الگوریتمی پیشنهاد گردید که هر دو ایده را در خود داشته باشد[1]. شکل ۲، روش پیشنهادی با ترکیب دو ایده را نشان می‌دهد. لازم به ذکر است که در این روش چهریزه‌های بنیادین شناسایی نمی‌شوند.

شکل ۲. الگوریتم پیشنهادی

مراحل اجرای الگوریتم پیشنهادی عبارت‌اند از:

گام اول. آماده‌سازی بدنهٔ متنی

با توجه به محدودیت زمانی انجام پژوهش، امکان اجرای روش پیشنهادی بر روی کلیهٔ حوزه‌های موضوعی وجود نداشت. از این رو، حوزهٔ پزشکی و دامنهٔ زنان و زایمان برای اجرای روش پیشنهادی در نظر گرفته شد. دلیل انتخاب حوزهٔ پزشکی:

◇ وجود رده‌بندی حوزهٔ پزشکی و مستقل و مجزا بودن دامنهٔ زنان و زایمان در

۱. مدل الگوهای تکراری بر اساس هم‌رخدادی واژگان پیشنهاد شده است که بیشتر برای علم‌سنجی استفاده می‌شود و در پایگاه‌های داده ساخت‌یافته نیز آزمون شده است و نه در محیط غیرساخت‌یافته، مانند این پژوهش. بنابراین، به نظر نگارنده، این ایده با شرایط این پژوهش همخوانی ندارد.

رده‌بندی «کتابخانهٔ پزشکی ملی آمریکا»، وجود اصطلاحنامهٔ پزشکی «مش»[1] که از اصول رده‌بندی منطقی و چهریزه‌ای تبعیت می‌کنند، و وجود اَبَراصلاحنامهٔ پزشکی یواِم‌ال‌اس»[2]؛

- ◇ وجود منابع دادهٔ رایگان در پایگاه «پاب‌مد»[3] برای مرور پشتوانهٔ انتشاراتی؛
- ◇ همچنین، عدم اجرای فعالیت مشابه در این دامنه.

حوزهٔ زنان و زایمان به درمان و پیشگیری بیماری‌های تولید مثل زنان و بیماری‌های پستان و عواملی می‌پردازد که بر سلامت زنان تأثیر می‌گذارد و همچنین به مراقبت و مداخلهٔ طبی و جراحی مناسب در حین بارداری و زایمان طبیعی و عارضه‌دار، مشاوره و آموزش در حاملگی، تنظیم خانواده و مراقبت از نوزاد مربوط است. با توجه به انتخاب پایگاه اطلاعاتی «پاب‌مد» برای کسب منابع داده‌ای رایگان جهت مرور پشتوانهٔ انتشاراتی، زبانِ اسناد مورد بررسی انگلیسی در نظر گرفته شد. برای تشخیص چهریزه‌ها در دامنهٔ تخصصی زنان و زایمان از پشتوانهٔ انتشاراتی ۲۰ مجلهٔ Q1 حوزهٔ زنان و زایمان در پایگاه JCR استفاده شد. سه مجموعهٔ متنی برای تحلیل مورد نیاز است که به‌گونه‌ای افراز از پشتوانهٔ انتشاراتی هستند:

۱. برای استخراج واژگان حوزه از یک نمونهٔ ۲۰۰۰ مقاله‌ای استفاده شد. به‌طور تصادفی از هر مجلهٔ Q1 حوزهٔ زنان و زایمان در پایگاه JCR، ۱۰۰مقاله در طول دوره ۲۰۱۴ تا ۲۰۱۸ از طریق پایگاه اطلاعاتی «پاب‌مد» استخراج شد. چکیده و عنوان هر مقاله در مجموعهٔ ۲۰۰۰ مقاله‌ای استخراج و به‌صورت یک فایل با فرمت txt. ذخیره‌سازی گردید. به‌دلیل اینکه این بدنهٔ متنی دارای واژگان نامناسب در تکرار بسیار زیاد مانند DOI، نام ناشر و مواردی از این دست است که سبب افزایش میزان خطا در زمان تحلیل می‌گردید، به‌صورت دستی اصلاح شدند.
۲. برای مجموعهٔ منشأ مقالات یک دورهٔ ۵ ساله از مجلات مذکور انتخاب شد. خروجی آن یک بدنهٔ متنی یکپارچه‌شده و اصلاح‌شده از پشتوانهٔ انتشاراتی در دورهٔ ۵ ساله بود که از این به بعد با نام مجموعهٔ منشأ[4] و با O نمایش داده می‌شود.
۳. برای مجموعهٔ بستر، تمامی مقالات در دورهٔ پنج‌ساله استخراج و با D نمایش داده می‌شود.

1. Mesh
2. Unified Medical Language System (UMLS)
3. Pubmed
4. original collection

گام دوم. استخراج واژگان مهم از بدنهٔ متنی

با توجه به بررسی مقالات مروری در حوزهٔ استخراج واژگان با تمرکز بر حوزهٔ پزشکی از مفاالـهٔ Zolfaghar et al. (2020) استفاده شد و برای تشخیص واژگان مهم در بدنهٔ متنی، از یک سرویس تحت وب با نام Biotex[1] که توسط آزمایشگاه هوش مصنوعی دانشگاه «مونت‌پولیه»[2] تهیه شده است، استفاده شد. این سرویس در سایت www.Tubo.lirmm.fr/biotex/ در دسترس است. این سرویس که بر اساس پژوهش‌های «ونچورا»[3] تهیه شده، دارای ۲۰۰ الگوی زبانی پزشکی است و همچنین امکان تطبیق واژگان با هستی‌شناسی «یواِم‌ال‌اس» یا زبان واحد پزشکی را داراست. این سرویس همچنین، امکان انتخاب معیار رتبه‌بندی واژگان را در اختیار کاربر می‌گذارد. معیارهای L-value، C-value، LIDF-value، F-OCapi، F-TFIDF-C، Okapi و TFIDF نمونهٔ معیارهای قابل انتخاب است. برای انتخاب معیار رتبه‌بندی از مقایسهٔ صورت‌گرفته توسط «ونچورا» که اشاره شد، استفاده شده است. در این مقایسه، معیار LIDF-value به‌عنوان معیار با دقت بیشتر در حوزهٔ پزشکی که نیاز به الگوهای زبانی خاصی دارد، ارزیابی گردیده است (Lossio-Ventura 2015; Lossio-Ventura et al. 2016).

ورودی این گام مجموعهٔ منشأ یا O است که در سرویس Biotex بارگذاری می‌شود. خروجی این سرویس یک فایل با فرمت XML است که فهرست رتبه‌بندی‌شدهٔ هر بدنهٔ متنی ورودی و تطبیق آن با هستی‌شناسی «یواِم‌ال‌اس» را دربرمی‌گیرد. خروجی این مرحله یک مجموعه از واژگان مهم در مجموعهٔ منشأ یا O است که با Ei={E1, E2, ..., En} نمایش داده می‌شود.

گام سوم. بسط واژگان در بافت موضوع

همان‌طور که در ایدهٔ اول اشاره شد، واژگان مهم در مجموعهٔ منشأ یا E_i باید توسط روشی در بافت یا بستر بسط داده می‌شدند تا واژه‌های مرتبط با آن‌ها شناسایی شوند. برای نمونه، واژهٔ «بارداری» یا pregnancy در مجموعهٔ E جزیی از چهریزهٔ «تولید مثل» یا reproduction است که در مجموعهٔ E دیده نمی‌شود. اما، در بافت دامنهٔ زنان و زایمان که در درخت «مش» شکل گرفته، دیده می‌شود. بنابراین، برای تشخیص چهریزه بایستی

1. biomedical term extraction (Biotex) 2. Montpellier 3. Ventura

واژه‌های مرتبط را در بافت شناسایی و سپس، چهریزه را تشخیص دهیم. از همین رو، برای توسعهٔ واژگان از درخت «مش» استفاده گردید. خروجی این گام فهرستی از واژگان مهم یا E و واژگان بسط یافته بر مبنای آن است که از این به بعد با نام D شناخته می‌شود و طبیعی است که E زیرمجموعهٔ D است.

گام چهارم. استخراج چهریزه‌های کاندید

در این گام چهریزه‌های کاندید بر اساس روش پیشنهادی «داکا» استخراج شدند. مراحل دستیابی به چهریزه‌های کاندید عبارت‌اند از:

در مجموعهٔ واژگان مهم یا E، به ازای هر t واژه تعداد اسنادی که واژه در آن رؤیت شده است، محاسبه و با df(t) نمایش داده می‌شود. واژه‌هایی که در مجموعهٔ E نیستند ولی در مجموعهٔ D قرار دارند، df(t) معادل با صفر در نظر گرفته می‌شود. واژگان بر اساس df(t) مرتب شده و رتبهٔ آن‌ها rank(t) نامیده می‌شود.

در مجموعهٔ واژگان توسعه‌یافته یا D، به ازای هر واژه t تعداد اسنادی که واژه در آن رؤیت شده، محاسبه و با $df_D(t)$ نمایش داده می‌شود. البته، این محاسبه در کلِ پایگاه اطلاعاتی «پاب‌مد» صورت می‌گیرد. همان‌طور که اشاره شد، واژه‌های مجموعهٔ E نیز در مجموعهٔ D قرار دارند و برای آن‌ها نیز بایستی محاسبه صورت پذیرد. واژگان بر اساس $df_D(t)$ مرتب شده و رتبهٔ آن‌ها با $rank_D(t)$ نامیده می‌شود.

برای هر واژه t در مجموعهٔ E و یا D، اگر دو شرط زیر برآورده شود، چهریزهٔ کاندید شناسایی می‌شود.

شرط ۱. انتقال مبتنی بر تکرار[1]: میزان اسناد مرتبط با واژهٔ t در بافت بیشتر از منشأ باشد یا به عبارتی

فرمول ۱ $$Shift_f(t)= df_D(t)- df(t)>0$$

مطابق با قانون «زیف»[2] برای توزیع تکرار واژگان، شرط فوق سبب می‌شود که واژگانی که در بافت نسبت به منشأ تکرار بالاتر دارند، در فهرست کاندیدها قرار گیرند.

شرط ۲. انتقال مبتنی بر رتبه[3]: رتبهٔ موجود واژه در بافت بایستی نسبت به منشأ رشد کند. این بدان معناست که واژگانی که در بافت افزایش رتبه دارند، به‌طور معمول، ویژگی

1. frequency-based shifting 2. Zipfian nature 3. rank-based shifting

بهتری برای عام شدن دارند. برای نمونه، واژهٔ «تولید مثل» در بافت عام‌تر از «بارداری» در منشأ است. پس، باید در رتبه‌های بالاتر قرار گیرد. به‌صورت تجربی نیز تأیید شده است که جست‌وجوی واژه‌های عام‌تر میزان تکرار سند بالاتری در پایگاه‌ها دارد. معادلهٔ زیر شرط دوم را نشان می‌دهد.

فرمول ۲

$$Shiftr(t)= \log2(rank(t)) - \log2(rankD(t))>0$$

خروجی این گام یک مجموعه از چهریزه‌های کاندید است.

پس از تعیین واژگان مهم بر اساس میزان تکرار در بافت، در این گام بایستی به‌صورت آماری معنادار بودن تمایز در تکرارها آزمون شود. مطابق با نظر «داکا» بهترین آمارهٔ درست‌نمایی که در شرایط مسئلهٔ فعلی قابل انجام است، استفاده از توزیع دوجمله‌ای است. به‌صورتی که

فرمول ۳

$$-\log\gamma_t = \log L(p_1, df_D(\mathrm{t}), |O|) + \log L(p_2, \mathrm{df(t)}, |O|) - \log L(p, df_D(\mathrm{t}), |O|) - \log L(p, \mathrm{df(t)}, |O|)$$

جایی که

$$L(p,k,n) = k\log(p) + (n-k)\log(1-p)$$

$$p_1 = \frac{df_D(\mathrm{t})}{|O|}$$

$$p_2 = \frac{\mathrm{df(t)}}{|O|}$$

$$p = \frac{p_1 + p_2}{2}$$

فهرست چهریزه‌های کاندید بر اساس $-\log\gamma_t$ به‌صورت افزایشی مرتب می‌شود. چهریزه‌های با رتبهٔ بالاتر از معناداری بیشتری نسبت به چهریزه‌های پایین‌تر برخوردار هستند. خروجی اصلی این گام، مجموعهٔ معتبر آماری چهریزه‌های کاندید است که با Facet(O) نام‌گذاری شده است.

گام پنجم. نهایی‌کردن چهریزه‌ها و اصلاح نام‌گذاری

مجموعهٔ چهریزه‌های رتبه‌بندی‌شده زمانی می‌توانند به‌عنوان چهریزهٔ نهایی انتخاب شوند که بتوان آن‌ها را در یک ساختار سلسله‌مراتبی جای داد و ارزش‌های[1] آن را شناسایی

1. value

کـرد و همچنیـن، از نام‌هـای شـناخته شـده اسـتفاده نمـود. بـه همیـن دلیـل، از ایـدهٔ دوم بـرای اصـلاح نام‌گـذاری و نهایی‌کـردن چهریزه‌هـا اسـتفاده شـد. مراحـل اجـرای ایـن گام عبارت‌انـد از:

۱. بـه ازای هـر واژه در مجموعـهٔ Facet(O) به‌صـورت دسـتی جایـگاه واژه در درخـت «مـش» شناسایی شد. بر اساس سه قاعده درخت هرس گردید:

قاعـدهٔ ۱. اخـص بـودن: اگـر واژه دارای میـزان کمتـر از ۵ بـرگ باشـد[1]، نشـان‌دهندهٔ اخص بـودن آن اسـت و بایسـتی حذف شـود.

قاعـدهٔ ۲. جایگزینـی: اگـر واژه بیـن ۵ تـا ۲۰ بـرگ بـوده و زیربـرگ داشـته باشـد، در ایـن صـورت، واژهٔ بالاتـر آن اسـتخراج و به‌عنـوان چهریـزه در نظـر گرفتـه می‌شـود.

قاعـدهٔ ۳. اعـم بـودن: اگـر واژه بیـش از ۲۰ بـرگ داشـته باشـد، به‌طـور مسـتقیم، به‌عنـوان چهریـزه شـناخته می‌شـود.

۲. نـام پیشـنهادی بـا اسـتفاده از سـرعنوان‌های «مـش» و گونه‌هـای معنایی«یـوامِ‌ال‌اس» و همچنیـن، روش‌هـای نام‌گـذاری چهریزه‌هـا انتخـاب گردیـد.

٤. یافته‌ها

بـرای اجـرای ایـن الگوریتـم نیـاز بـه سـه مجموعـهٔ متنـی از درون پشـتوانهٔ انتشـاراتی دامنـهٔ تخصصـی زنـان و زایمـان اسـت. بـرای ایجـاد مجموعـهٔ متنـی بسـتر (حـوزهٔ زنـان و زایمـان) از چکیـده و عنـوان مجموعـهٔ مقـالات موجـود در ۲۰ مجلـهٔ برتـر حـوزهٔ زنـان و زایمـان اسـتفاده شـده اسـت. ایـن مجموعـه دربرگیرنـدهٔ ۱۶۷۰۷۱ سـند در زمـان نوشـتار حاضـر اسـت. بـرای ایجـاد بدنـهٔ منشـأ از بیـن ۱۶۷۰۷۱ سـند، اسـناد مربـوط بـه ۵ سـال اخیـر کـه معـادل ۲۶۴۳۰ سـند اسـت، اسـتخراج گردیـد. بـرای شناسـایی واژه‌هـای مهـم در بسـتر اصلـی از نمونه‌گیـری ۲۰۰۰ مقالـه کـه به‌صـورت تصادفـی انتخـاب شـده‌اند، اسـتفاده شـد. عنـوان و چکیـدهٔ ایـن مقـالات به‌صـورت فایـل txt. اسـتخراج شـده و جهـت ورود بـه سـرویس اسـتخراج واژگان مهـم اصـلاح شـدند. برخـی از واژه‌هـای بسـیار تکـراری ماننـد DOI و یـا ناشـر و عنـوان مجلـه حـذف شـده و به‌دلیـل اینکـه از روش LIDF-value بـرای رتبه‌بنـدی واژگان مهـم اسـتفاده می‌شـد، تقسـیم‌کننده در بیـن

---

۱. میـزان عـدد ۵ و ۲۰ وابسـته بـه تواتـر برگ‌هـا در یـک درخـت اسـت کـه بـرای درخـت «مـش» و در ایـن پژوهـش شناسـایی شـده اسـت. طبیعـی اسـت کـه در پژوهـش دیگـر، دامنـهٔ دیگـر ایـن اعـداد متفـاوت خواهنـد بـود. تنظیـم ایـن پارامتـر تـا حـدود زیـادی تجربـی اسـت.

هر سند اضافه گردید تا امکان اجرای روش مهیا گردد. برای شناسایی واژگان مهم از سرویس BioTex استفاده گردید و ورودی متنی ایجادشده با حجم ۴ مگابایت که معادل ۵۸۱۷۰۰ واژه بود، در آن بارگزاری شد.

خروجی حاصل از سرویس BioTex دربرگیرندهٔ ۱۲۰۰ واژهٔ مهم استخراج‌شده از متون است که به‌صورت صعودی ارائه شده است. این سرویس برای شناسایی واژگان معتبر در حوزهٔ پزشکی، واژگان را با واژگان هستی‌شناسی «یوام‌الاس» مطابقت داده و فهرست معتبری از واژگان را ارائه می‌دهد. از میان ۱۲۰۰ واژه، ۵۱۴ واژه به‌عنوان واژگان معتبر شناسایی شد. با بررسی واژگان مشخص شد که برخی از واژگان تکراری هستند؛ برای نمونه، Cell و Cells هر دو جزء واژه‌های مهم بودند که یکی از آن‌ها در نظر گرفته شد و در نهایت، ۴۸۰ واژه به‌عنوان واژهٔ مهم منظور شد. جدول ۲، نمونه‌ای از واژگان و رتبه‌بندی آن‌ها را نشان می‌دهد.

جدول ۲. نمونه‌ای از فهرست واژگان مهم

ردیف	واژه	ردیف	واژه
۱	breast cancer	۱۱	pelvic floor
۲	cervical cancer	۱۲	ovarian tissue
۳	preterm birth	۱۳	cesarean delivery
۴	endometrial cancer	۱۴	growth factor
۵	embryo transfer	۱۵	birth weight
۶	task force	۱۶	pregnancy loss
۷	live birth	۱۷	oxidative stress
۸	pregnancy rate	۱۸	first trimester
۹	vitamin d	۱۹	blood loss
۱۰	Body Mass Index	۲۰	urinary incontinence
	.....		.....

پس از شناسایی واژگان مهم، مطابق الگوریتم بایستی یک مجموعه از واژگان توسعه یافته ایجاد شود که با استفاده از ساختار درختی «مش» گسترش یافته است. هر واژه در

درخت «مش» شناسایی گردیده و سطوح بالاتر آن به‌عنوان واژه‌های گسترش یافته ارائه شده‌اند. برای نمونه، این واژه در دو بخشِ درخت رؤیت شده است که عبارت‌اند از:

Urogenital System [A05] ← **development word**

Genitalia [A05.360]

Genitalia, Female [A05.360.319]

Genitalia, Male [A05.360.444]

Germ Cells [A05.360.490] ← **important word**

Ovum [A05.360.490.690]

Spermatozoa [A05.360.490.890]

Gonads [A05.360.576]

که واژهٔ سرمنشأ آن urogential system است که به‌عنوان یک واژهٔ گسترش‌یافته در نظر گرفته می‌شود. شاخهٔ دوم نیز به‌صورت زیر است.

Cells [A11] ← **development word**

Erythroid Cells [A11.443]

Eukaryotic Cells [A11.450]

Germ Cells [A11.497] ← **important word**

Embryonic Germ Cells [A11.497.124]

Germ Cells, Plant [A11.497.248]

Ovum [A11.497.497]

.....

که سرمنشأ آن cells است که به‌عنوان توسعه‌یافته در نظر گرفته می‌شود. البته، واژهٔ cells افزون بر توسعه، در واژگان مهم نیز قرار دارد که در این صورت، یکی از آن‌ها در

نظـر گرفتـه می‌شـود. پـس از اجـرای توسـعهٔ واژگان ۶۶۶ واژه نهایـی شـد و به‌عنـوان واژگان توسـعه‌یافته در نظـر گرفتـه شـد.

مطابـق بـا الگوریتـم پیشـنهادی، پـس از شناسـایی واژگان مهـم در منشـأ و بسـتر نیـاز اسـت کـه تعـداد اسـنادِ دربرگیرنـدهٔ واژه در منشـأ و بسـتر محاسـبه گـردد. بـرای ایـن امـر از یـک اسـتراتژی جسـت‌وجو اسـتفاده شـده اسـت. در ایـن اسـتراتژی کـه به‌صـورت زیـر صورت‌بنـدی شـده، ابتـدا تعـداد سـند در منشـأ و سـپس، در بسـتر محاسـبه می‌شـود.

```
(((((((((((((((( breast cancer"[Title/Abstract]) AND "Hum
Reprod Update"[Journal]) OR "Am J Obstet
Gynecol"[Journal]) OR "Obstet Gynecol"[Journal])
OR "BJOG"[Journal]) OR "Hum Reprod"[Journal]) OR
"Gynecol Oncol"[Journal]) OR "Ultrasound Obstet
Gynecol"[Journal]) OR "Fertil Steril"[Journal]) OR
"Pregnancy Hypertens"[Journal]) OR "Mol Hum
Reprod"[Journal]) OR "Maturitas"[Journal]) OR
"Reprod Biomed Online"[Journal]) OR "Clin
Perinatol"[Journal]) OR "Semin
Perinatol"[Journal]) OR "J Gynecol
Oncol"[Journal]) OR "J Minim Invasive
Gynecol"[Journal]) OR "Contraception"[Journal])
OR "Breast"[Journal]) OR "Placenta"[Journal]) OR
"Best Pract Res Clin Obstet Gynaecol"[Journal] AND
“breast cancer"[Title/Abstract])
```

شکل ۳. راهبرد جست‌وجو در منشأ (۲۶۴۳۰ سند) و بستر (۱۶۷۰۷۱سند)

پـس از اسـتخراج اعـداد آن‌هـا بـر اسـاس صعـودی بـه نزولـی رتبه‌بنـدی می‌شـوند. جـدول ۳، نمونـه‌ای از محاسـبات را نشـان می‌دهـد.

جدول ۳. نمونه‌ای از محاسبهٔ تعداد اسناد دربرگیرندهٔ واژه در منشأ و بستر

واژه های مهم در منشأ	تعداد اسناد دربرگیرندهٔ واژه در منشأ	رتبه در منشأ	تعداد اسناد دربرگیرندهٔ واژه در بستر	رتبه در بستر	مقایسه براساس فرمول ۱ و ۲	
	df(t)	rank(t)	$df_D(t)$	$rank_D(t)$	شرط ۱	شرط ۲
5-hydroxyvitamin d	۴۷	۱۵۹	۸۴	۳۸۷	✓	
Abdomen (facet)	۰	۲۰۵	۶۸۸	۱۷۲	✓	✓

واژه های مهم در منشأ	تعداد اسناد دربرگیرندهٔ واژه در منشأ	رتبه در منشأ	تعداد اسناد دربرگیرندهٔ واژه در بستر	رتبه در بستر	مقایسه براساس فرمول ۱ و ۲	
Abdominal Cavity	۰	۲۰۵	۱۶۵	۳۳۲	✓	
abdominal circumference	۷۲	۱۳۵	۳۷۸	۲۴۴	✓	
abdominal hysterectomy	۸۰	۱۲۸	۸۲۵	۱۵۰	✓	
Abdominal Muscles	۰	۲۰۵	۳	۴۵۵	✓	
abdominal wall	۶۱	۱۴۵	۴۰۸	۲۳۱	✓	
Abortion (facet)	۰	۲۰۵	۴۶۸۰	۳۳	✓	✓
absolute risk	۵۴	۱۵۲	۱۱۸	۳۶۱	✓	
activin a	۲۲	۱۸۳	۱۳۸	۳۴۷	✓	
.....	.....	.....	.....	.....		

مطابق با فرمول شماره ۱ و ۲ مقایسه بین رتبه‌ها و تعداد اسناد صورت‌گرفته و چهریزه‌های کاندید شناسایی شدند که در جدول ۴، قابل مشاهده است. برای بررسی اعتبار چهریزه‌ها از فرمول ۳، استفاده شد. نتایج این ارزیابی به‌صورت صعودی مرتب شده و چهریزه‌ها شناسایی می‌شوند. جدول ۴، بخشی از چهریزه‌ها و میزان اعتبار آن‌ها را نشان می‌دهد.

جدول ٤. نمونه‌ای از چهریزه‌ها و رتبه‌بندی آن بر اساس تابع توزیع دوجمله‌ای

واژه‌های مهم در منشأ	تعداد اسناد دربرگیرندهٔ واژه در منشأ $df(t)$	تعداد اسناد دربرگیرندهٔ واژه در بستر $df_D(t)$	$-\log \gamma_t$
Pregnancy	۱	۲۶۴۲۹	۱۵۹۰۲/۷۳
Risk	۱	۲۴۷۵۲	۸۲۶۷/۴۹
Disease	۱	۱۶۷۷۸	۵۴۱۱/۷۸
Blood	۱	۱۴۶۲۷	۴۶۷۰/۸۲
Diagnosis	۱	۱۴۵۵۳	۴۶۴۵/۷۰
Therapy	۱	۱۴۳۲۸	۴۵۶۹/۳۹
Cells	۱	۱۳۵۸۷	۴۴۱۰/۱۳

واژه‌های مهم در منشأ	تعداد اسناد دربرگیرندهٔ واژه در منشأ df(t)	تعداد اسناد دربرگیرندهٔ واژه در بستر $df_D(t)$	$-\log \gamma_t$
surgery	۱	۱۲۱۵۲	۳۸۳۸/۹۶
Research	۱	۱۱۰۱۲	۳۴۶۱/۶۵
Health	۱	۱۰۰۲۸	۳۱۳۸/۸۸
Infertility	۱	۸۹۸۶	۲۷۹۹/۹۸
Tissue	۱	۸۶۳۲	۲۶۸۵/۵۲
Fertilization	۱	۸۰۰۸	۲۴۸۴/۵۷
.....	....	.....	.....

اما در گام آخر، برای تشخیص چهریزه‌های نهایی نیاز است که هر واژه که به‌عنوان چهریزه در نظر گرفته شده در درخت «مش» جست‌وجو شود و سپس، بر اساس قواعد ذکرشده در الگوریتم، چهریزه‌های نهایی پیشنهاد شود. جدول ۵، نمونه‌ای از این قواعد را نشان می‌دهد.

جدول ۵. نمونه‌ای از اصلاح چهریزه‌های پیشنهادی بر مبنای قواعد هرس درخت «مش»

چهریزهٔ پیشنهادی	تعداد برگ زیر دست در درخت	مقایسه	اعمال قانون
Pregnancy	۱۶	>۵	اول (جایگزینی)= جایگزینی واژهٔ بالادست گره یعنی (reproduction)
Therapy	>۲۰	>۲۰	دوم (اعم بودن)= واژه چهریزه است.
Cells	>۲۰	>۲۰	دوم (اعم بودن)= واژه چهریزه است.
Placenta	۳	<۵	سوم (اخص بودن)= حذف چهریزه

پس از اعمال قواعد بر روی چهریزه‌ها مجموعه‌ای از چهریزه‌های پیشنهادی به‌عنوان چهریزهٔ نهایی در نظر گرفته شدند که در جدول ۶، فهرست شده‌اند.

جدول ٦. فهرست چهریزه‌های نهایی

ردیف	چهریزه‌های نهایی	ردیف	چهریزه‌های نهایی
۱	Reproduction	۱۴	Age Groups
۲	risk factor	۱۵	reproductive physiological phenomena
۳	pathologic process or disease	۱۶	Hormones

ردیف	چهریزه‌های نهایی	ردیف	چهریزه‌های نهایی
۴	Diagnosis	۱۷	Prognosis
۵	Therapy	۱۸	diagnostic techniques
۶	Cells	۱۹	Education
۷	Research	۲۰	Steroids
۸	Health	۲۱	Epidemiology
۹	Tissue	۲۲	Lipids
۱۰	Signs and Symptoms	۲۳	body regions
۱۱	ebryonic structures	۲۴	genecologic Surgical Procedures
۱۲	female gentalia	۲۵	Persons
۱۳	reproductive techniques	۲۶	peregnancy complication

۵. بحث و نتیجه‌گیری

در بخش اولِ بحث و نتیجه‌گیری، الگوریتم پیشنهادی با الگوریتم‌های ارائه‌شده در پژوهش پیشین مقایسه شده است. از نگاه رایانه‌ای، الگوریتم پیشنهادی به مسئلهٔ بهینه‌سازی سنتی شباهت دارد؛ یعنی چهریزه‌ها بر اساس چند معیار انتخاب می‌شوند. اما با پیچیده‌شدن نیازهای کاربران مانند شناخت بافت متون و درک رابطهٔ چند مفهوم که در جست‌وجوی اولیه برای کاربر نامشهود است، نیاز به بازنگری در مسئلهٔ بهینه‌سازی ضرورت یافت. برای دستیابی به هدف «دسترسی اکتشافی فعالانه» نیاز به فرمول‌بندی و توصیف مسئله به‌گونه‌ای دیگر است. در این صورت، فرمول‌بندی مسئله دارای شرایط زیر است:

۱. جست‌وجوی اولیه بر اساس کلیدواژه (حل مسئلهٔ بهینه‌سازی سنتی)؛
۲. شناخت روابط میان واژهٔ جست‌وجوشده و دیگر واژگان (نیاز به یک مجموعه واژگان کنترل‌شده و مرتبط، مانند اصطلاح‌نامهٔ چهریزه‌ای یا هستی‌شناسی) در بافت موضوع (چهریزه‌های در حوزهٔ علمی یا چهریزه‌های در دامنه‌های فرعی در حوزه یا استفاده از گونه‌های معنایی در هستی‌شناسی)؛
۳. ایجاد پرس‌وجوهای تودرتو بر اساس پردازش مرحلهٔ ۲.

تنها تفاوت طرح الگوریتم پیشنهادی با طرح مسئلهٔ سنتی در این است که سیستم

بر اساس یک مجموعه مربی[1] امکان درک بیشتر یا یادگیری فعالانه را در کاربر تقویت می‌کند و قاعدتاً به نتایج مرتبط‌تر با نیاز اطلاعاتی کاربر می‌رسد. اما پرسش اصلی این است که در هر دامنهٔ تخصصی چه چهریزه‌هایی در نظر گرفته شود. با توجه به اینکه، هدف اساسی در استفاده از سیستم‌های رایانه‌ای غلبه بر میزان بسیار بالای اطلاعات در پایگاه‌های اطلاعاتی امروزی است، بنابراین، شناسایی چهریزه‌ها نیز بایستی از یک مکانیزم خودکار برخودار باشد. این است که از الگوریتم استخراج خودکار پیشنهادی استفاده شد. جدول ۷، مقایسهٔ الگوریتم پیشنهادی با دیگر الگوریتم‌ها را نشان می‌دهد.

جدول ۷. مقایسهٔ الگوریتم پیشنهادی با الگوریتم‌های دیگر

الگوریتم‌ها	ورودی	الگوریتم انتخاب واژگان	روش کار	مربی
Castanet (Stoica, Hearst and Richardson 2007)	بدنهٔ متنی	توزیع آماری واژگان	ایجاد درخت واژگان و خوشه‌بندی آن	«وردنت»
Daka (Dakka 2008; Dakka, Ipeirotis and Sacco 2009)	بدنهٔ متنی	بسامد تکرار	ایجاد دو افراز (افراز منشأ و بدنهٔ متنی) و مقایسهٔ رفتار واژه در آن‌ها	–
Dynacet (Basu Roy 2011)	پایگاه اطلاعاتی	–	کدام واژه میزان پرس‌وجو را کمینه می‌کند؟	–
Facetpedia (Li et al. 2010)	بدنهٔ متنی	–	برای هر واژه یک درخت ایجاد می‌شود و سپس، با خوشه‌بندی درخت چهریزه‌ها شناسایی می‌شوند.	«ویکی‌پدیا»
پیشنهادی پژوهش حاضر	بدنهٔ متنی	LIDF-Value	ایجاد سه افراز (افراز منشأ و بدنهٔ متنی و افراز برای شناسایی واژگان مهم) و مقایسهٔ رفتار واژه در آن‌ها و سپس، ایجاد درخت بر اساس چهریزه‌های کاندید و سپس، خوشه‌بندی آن‌ها و استخراج چهریزهٔ نهایی	«مش»

همان‌طور که در جدول بالا اشاره شد، الگوریتم پیشنهادی سعی کرده است با ترکیب دو الگوریتم، ضعف حاصل از دیگر الگوریتم‌ها را بهبود دهد. در الگوریتم «داکا» که به‌عنوان الگوریتم پایه در نظر گرفته شد، نتایج حاصل موارد زاید زیادی دارند؛ یعنی

1. supervisor

واژگانی به‌عنوان چهریزه انتخاب می‌شوند که به‌صورتی در دیگر چهریزه‌ها قابل ادغام هستند. برای نمونه واژهٔ abortion یا «سقط جنین» بر اساس اصول چهریزه‌ای نمی‌تواند چهریزه باشد؛ چرا که نمی‌توان بر اساس «سقط جنین» تقسیم‌بندی اعمال نمود. دلیل «سقط جنین» در علائم و نشانه‌ها یا بیماری‌ها جای می‌گیرد. بنابراین، از یک روش درختی با الگوگیری از الگوریتم Castanet استفاده شد. برای این کار نیاز به مجموعه مربی بود که «مش» در نظر گرفته شد. این ادغام سبب شد که نتایج حاصل از الگوریتم «داکا» اصلاح شود و چهریزه‌هایی پیشنهاد شوند که نه خیلی عام و نه خیلی خاص باشند.

پس از بررسی الگوریتم پیشنهادی، بایستی به این موضوع توجه شود که نتایج این الگوریتم یعنی چهریزه‌های انتخاب‌شده چه تفاوت و شباهتی با دیگر چهریزه‌های سنتی دارند. در سنت چهریزه‌ای، رده‌بندی «کولُن» و رده‌بندی «بلیس ۲» در حوزهٔ پزشکی، چهریزه‌های پیشنهادی خود را ارئه نموده‌اند. «رانگاناتان» برای حوزهٔ پزشکی چهار چهریزه ارائه نموده است که به‌گونه‌ای به چهریزه‌های بنیادین متصل است[1] (Ranganathan 1933). از سوی دیگر، در رده‌بندی «بلیس ۲» در ردهٔ پزشکی، ده چهریزه ارائه شده است[2]. در کل، چهریزه‌های پزشکی «رانگاناتان» و «بلیس» تا حدودی نشان‌دهندهٔ موضوعات اساسی هستند، ولی برای استخراج نهایی چهریزه‌های حوزهٔ پزشکی می‌توان به زبان واحد پزشکی یا «یوامال‌اس» رجوع نمود. گونه‌های معنایی ارائه‌شده توسط «یوامال‌اس» از دیدگاه متخصصان، از جمله Broughton (2008) همان چهریزه‌ها هستند و اگر گونه‌های معنایی «یوامال‌اس» با چهریزه‌های پیشنهادی «رانگاناتان» و «بلیس» نیز مقایسه شوند، بسیار شباهت دارند. برای بررسی و تحلیل نتایج، چهریزه‌های مشترک بین «رده‌بندی کولُن»، «رده‌بندی بلیس» و «یوامال‌اس» شناسایی و با چهریزه‌های شناسایی‌شده در پژوهش حاضر مقایسه شد. جدول ۸، این مقایسه را نشان می‌دهد.

1. problem, organ, cause (for disease) of [energy], handling (for disease) of [energy]
2. types of mankind, race, processes of the human organism, individuals, processes, parts, organs, systems actions on persons, agents of actions, common facets

جدول ۸. مقایسهٔ چهریزهٔ پیشنهادی با چهریزه‌های رده‌بندی «کولُن»، «بلیس» و «یوام‌ال‌اس»

چهریزه‌های بنیادی	چهریزه‌های حوزهٔ پزشکی	توضیحات	چهریزه‌های شناسایی‌شدهٔ رایانه‌ای (پژوهش حاضر)
entity	organism	جاندار یا اَندامگان یا ارگانیسم به معنای یک سامانهٔ زنده و پیچیده از اعضاست که با تأثیرشان بر یکدیگر، امکان سازگاری با محیط و تضمین بقا و پایداری کلّ آن موجود زنده را فراهم می‌کنند. اَندامگان‌ها انواع متنوعی دارند؛ به‌عنوان مثال، شامل تک‌یاختگان، ویروس‌ها، قارچ‌ها، گیاهان و جانوران می‌شوند. انسان یک گونه از جانوران است که بر اساس ویژگی‌های جنسیت، سن، نژاد و موارد شبیه آن قابل تقسیم‌بندی است.	age groups body regions
part	anatomy	کالبدشناسی یا علوم تشریح یا آناتومی شاخه‌ای از زیست‌شناسی است که به بررسی ساختار و شیوهٔ کار اجزای ارگانیسم‌ها می‌پردازد.	tissue female gentalia cells
process	process	منظور، فرایندهایی زیستی است که در حین تکامل یک ارگانیسم رخ می‌دهند و سبب رشد یا یک ناهنجاری یا بیماری در یک ارگان یا مجموعهٔ ارگان‌ها می‌گردند. همچنین، فرایندهای غیرزیستی (انسانی) مانند تصادفات یا مسمومیت‌هاست که بر رشد، ناهنجاری و بیماری‌های ارگان‌ها یا مجموعهٔ ارگان‌ها اثر می‌گذارد.	reproductive physiological phenomena ebryonic structures
operation	operation	منظور، مجموعه فعالیت‌های آزمایشی بررسی‌های بالینی، شرح تاریخی اقدامات بر روی یک کارکرد ارگانیسم انسانی است که امکان تشخیص را فراهم می‌کند. همچنین، فعالیت‌های درمانی، تحقیقاتی و آموزشی را دربرمی‌گیرد.	peregnancy complication genecologic Surgical Procedures education diagnostic techniques prognosis epidemiology reproductive techniques signs and symptoms health research therapy diagnosis
agent	agent	منظور، مجموعه پرسنل بهداشتی، ابزارها و داروهایی است که فعالیت‌های تشخیصی، درمانی، تحقیقاتی، آموزشی و آزمایشگاهی را انجام می‌دهند.	persons

چهریزه‌های بنیادی	چهریزه‌های حوزهٔ پزشکی	توضیحات	چهریزه‌های شناسایی‌شدهٔ رایانه‌ای (پژوهش حاضر)
material	substance	مواد شیمیایی، ترشحاتی و غذا به‌عنوان مادهٔ ایجادکنندهٔ انرژی ارگانیسم انسانی	lipids hormones steroids
space	space	چهریزهٔ عمومی	-
time	time	چهریزهٔ عمومی	-

با توجه به مقایسهٔ صورت‌گرفته در جدول ۸، موارد زیر قابل توجه است.

۱. چهریزه‌های شناسایی‌شده در حوزهٔ زنان و زایمان خُرد هستند. برای نمونه، چهریزهٔ سلول (cells) یا بافت (tissue) جزیی از چهریزهٔ آناتومی هستند. این بدان معناست که چهریزه‌های عمومی در «رده‌بندی کولُن»، «رده‌بندی بلیس» در حوزهٔ پزشکی بسیار عام هستند و قابلیت مرور بر اساس چهریزه‌ها را ایجاد نمی‌کنند؛ در حالی که رویکرد رایانه‌ای به چهریزه‌های خُردتر اشاره دارد، زیرا که در بافت موضوعی تمرکز می‌کند. تنها در فرااصطلاحنامه «یواِم‌ال‌اس» افزون بر چهریزه‌های عام، چهریزه‌های خُرد هم وجود دارند.

۲. برخی از چهریزه‌های استخراج‌شده به‌صورت خودکار بر اساس الگوریتم پیشنهادی، بسیار خُرد هستند. برای نمونه، اندام جنسی female gentalia در سطحی قرار گرفته است که امکان ایجاد مرور برای آن قابل درک نیست. البته، تعداد این چهریزه‌ها تنها دو مورد است که female gentalia و embryonic structure است و دلیل آن خطای الگوریتم است.

۳. چهریزه‌های عمومی در روش پیشنهادی پژوهش حاضر شناسایی نشده‌اند. این امر به این دلیل است که واژه‌های مکان یا زمان تواتر کمتری نسبت به واژه‌های تخصصی دارند و در اولویت واژگان قرار نگرفتند. به نظر می‌رسد که اطلاعات کتابشناختی منابع باید جداگانه بررسی گردد تا چهریزه‌های عمومی شناسایی شوند.

مقایسهٔ چهریزه‌های استخراج‌شده در این پژوهش با چهریزه‌های استخراج‌شده با رویکرد کاربرمدار در حوزهٔ زنان و زایمان (Farajpahlou et al. 2020) نشان می‌دهد که حدود نیمی از چهریزه‌های پیشنهادی در هر دو رویکرد همپوشانی دارند. دلیل تفاوت این است که در رویکرد کاربرمدار، بر اساس پشتوانهٔ دانش کاربران، پاسخ‌های داده‌شده به‌طور غیرمستقیم طبقه‌بندی تعلیمی را آشکار می‌کنند. اما رویکرد محاسباتی، بر اساس

پشتوانهٔ انتشاراتی، چهریزه‌های استخراج‌شدهٔ وابسته به واژگان مستخرج از متون را ارائه می‌دهد. نکتهٔ دیگر اینکه، مجموعهٔ چهریزه‌های ایجادشده در هر دو رویکرد، طبقه‌بندی بنیادین را حاصل نکرده است. دلیل این امر می‌تواند به این امر بازگردد که طبقهٔ بنیادین یا چهریزه‌های بنیادین حالت تجریدی محض دارند و به‌طور عمده توسط متخصصان حوزهٔ رده‌بندی و نه متخصصان حوزهٔ موضوعی، شکل گرفته‌اند.

یکی از پرسش‌های این پژوهش به بررسی وابستگی یا عدم وابستگی چهریزه‌ها به دامنهٔ تخصصی بازمی‌گردد. در این پژوهش، چهریزه‌های مختص دامنهٔ زنان و زایمان استخراج شد و نتایج نشان داد که چهریزه‌های پیشنهادی، خُردتر و برای مرور مناسب‌تر هستند. با توجه به اینکه در پژوهش‌های گذشته نیز به وابستگی یا استقلال چهریزه‌ها به دامنهٔ تخصصی توجه نشده، برای پاسخ دقیق به این سؤال بایستی چندین دامنهٔ دیگر نیز مورد بررسی قرار گیرند. در حقیقت، بایستی چنین پژوهشی را در دامنه‌های دیگر انجام گیرد تا با مقایسهٔ آن‌ها پاسخی مناسب به این پرسش داده شود. اما به‌طور کلی، می‌توان به موارد زیر اشاره کرد:

۱. در این پژوهش مشخص شد که چهریزه‌های دامنهٔ تخصصی از چهریزه‌های عمومی در حوزهٔ پزشکی که در «رده‌بندیکولُن»، «رده‌بندی بلیس» و «یوام‌الاس» ارائه شده‌اند، متفاوت بوده و مستقل از آن‌ها قابل شناسایی و تعریف هستند.
۲. الگوریتم پیشنهادی بایستی بتواند بافت را تشخیص دهد. بدین معنا که چهریزه‌های ارائه‌شده بسیار کلی نباشند که کاربرد عام داشته باشند. چهریزه‌ها باید بتوانند حوزهٔ اصلی (برای نمونهٔ پزشکی) و دامنهٔ تخصصی (زنان و زایمان) را نشان دهند. بدین معنا که اگر شخصی چهریزه‌ها را در بازیابی اطلاعات برای محدود کردن جست‌وجو مشاهده کرد، متوجه دامنهٔ تخصصی آن بشود. اما نکته این است که به‌جز معدودی از چهریزه‌ها، بقیهٔ موارد مانند تشخیص، درمان، سلول، هورمون، پیش‌آگهی و ... در دیگر دامنه‌های پزشکی نیز قابل شناسایی هستند. بنابراین، از این نظر نیاز به پژوهش بیشتری است تا در چند دامنهٔ دیگر بررسی صورت گیرد و در صورت عدم تغییر در چهریزه‌ها، می‌توان به عدم استقلال چهریزه‌ها در دامنه‌های مختلف اشاره کرد.

دو محدودیت بر روند اجرای پژوهش حاضر حاکم بوده است که عبارت‌اند از:

۱. دقت روش محاسباتی به میزان حجم بدنهٔ متنی بازمی‌گردد. برای نمونه، اگر از بدنهٔ متنی تمام‌متن در یک دورهٔ ده‌ساله استفاده می‌شد، قاعدتاً این تفاوت‌ها رفع

می‌گردید. اما حجم این بدنهٔ متنی با توجه به یک برداشت توسط محقق ۱/۳ گیگابایت است که امکان پردازش آن در هیچ پردازشگری وجود ندارد. بالاترین حجم پردازش در سیستم مورد استفادهٔ این پژوهش ۶ مگابایت است که فاصلهٔ بسیار زیادی با حجم بدنهٔ متنی تمام‌متن دارد.

۲. در روش رایانه‌ای، هر سند یک مجموعه ویژگی‌های عمومی دارد؛ مانند نام نویسنده، زمان انتشار، نام مجله، نوع سند، زبان و .... که در بدنهٔ متنی که ویژگی‌های خاص سند است، قرار ندارند. به همین دلیل، الگوریتم رایانه‌ای آن‌ها را تشخیص نمی‌دهد.

طبیعی است که در حین اجرای پژوهش مسائلی شناسایی می‌شوند که پژوهش دربارهٔ آن‌ها می‌تواند پیشنهادهایی برای پژوهش‌های آینده باشند. از این رو، فهرستی از پیشنهادهای پژوهشی در ادامه ذکر شده است:

- ◇ از منظر ارائهٔ الگوریتم‌ها برای جست‌وجوی چهریزه‌ای، الگوریتم‌های هوشِ مصنوعی مانند الگوریتم ژنتیک، الگوریتم‌های تکاملی و فازی برای استخراج چهریزه‌ها مورد بررسی و پژوهش قرار نگرفته‌اند. همچنین، تمامی پژوهش‌ها تاکنون به‌گونه‌ای به ارائهٔ روش‌های با سرپرست (استفاده از اصطلاحنامه یا هستی‌شناسی) برای استخراج چهریزه‌ها پرداخته‌اند. پیشنهاد می‌شود در پژوهش‌های آینده به ارائهٔ روش‌های بدون سرپرست توجه شود.
- ◇ برای استخراج چهریزه‌ها پیشنهاد می‌شود که روش‌های دیگر متن‌کاوی مانند استفاده از هم‌رخدادی واژگان در استخراج چهریزه‌ها مورد بررسی و پژوهش قرار گیرد. بدین‌صورت که پژوهشی تنظیم شود که از هم‌رخدادی واژگان و خوشه‌بندی آن به مجموعه‌ای از چهریزه‌ها (نه موضوعات) دست یابد.
- ◇ همان‌طور که اشاره شد، یکی از محدودیت‌های روش‌های محاسباتی برای شناسایی چهریزه‌ها، هزینهٔ پردازشی بالای فعالیت‌ها در داده‌های حجیم است که می‌تواند به‌عنوان یک موضوع مستقل یعنی «استخراج چهریزه‌ها در داده‌های حجیم» پیشنهاد گردد.

References

Basu Roy, S. 2011. Efficient exploration techniques on large databases. PhD diss., University of Texas, Arlington.

Bliss, H. E. 1940-1953. *A Bibliographic Classification, Extended by Auxiliary Schedules for Composite Specification and Notation*. (4 vols. in 3):Vol. 1 (Classes 1–9, A–G) 1940; Vol. 2 (Classes H–K) 1947 (a 2nd edition of Vols. 1 and 2 appeared in 1952, in one volume); Vol. 3 (Classes L–Z), and Vol. 4 (General Index). NewYork: Wilson.

Dakka, W. 2008. Faceted searching and browsing over large collections of textual and text-annotated objects. PhD diss., Columbia University.

_____, P. Ipeirotis, & G. M. Sacco. 2009. *Taxonomy Design Dynamic Taxonomies and Faceted Search*. NewYork: Springer.

Farajpahlou, A., F. Osareh, S. M. Fakhrahmad, & L. Dehghani. 2019. The development of facet analysis approach in knowledge organization: A 100-year review. *Iranian Journal of Information processing and Management* 34 (3): 1235-64.

_____. 2020. The User-Oriented Approach for Facet Extraction in Gynecology and Obstetrics Domain. *Health Information Management* 16 (6): 285-293.

Hearst, M. A. 2006. Clustering versus faceted categories for information exploration. *Communications of the ACM* 49 (4): 59-61.

Hjørland, B. 2003. Fundamentals of knowledge organization. *Knowledge Organization* 30 (2): 87-111.

_____. 2013. Facet analysis: The logical approach to knowledge organization. *Information Processing & Management* 49 (2)545-557 :.

Hudon, M. 2020. Facet. *Knowledge Organization* 47 (4):.320-333

Komamizu, T. A. 2015. Study on Faceted Search for Semi-structured Data. PhD diss., University of Tsukuba.

Li, C., N. Yan, S. B. Roy, L. Lisham, and G. Das2010 .. Facetedpedia: dynamic generation of query-dependent faceted interfaces for wikipedia. Paper presented at the Proceedings of the 19th International Conference on World Wide Web. North Carolina, USA: 456-499.

Lossio-Ventura, J. A. 2015. Towards the French biomedical ontology enrichment. PhD diss. University of Montpellier.

Lossio-Ventura, Juan Antonio, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire 2016. Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal* 19 (1-2):59-99 .

Mauro, N., G. Izzi, M. Pellegrino, L. Ardissono, C. Grandi, M. Lucenteforte, & M. Segnan. 2020. Faceted Exploration of Cultural Heritage. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*340-346 :. Genoa Italy340-346 :.

Ranganathan, S. R. 1933. *Colon clasification*. Madras, India: Madras Library Association.

Sacco, G. M. 2000. Dynamic taxonomies: A model for large information bases. *Knowledge and Data Engineering*, IEEE Transactions on, 12 (3).468-479 :

Sacco, G. M., Y. Tzitzikas. 2009. *Dynamic Taxonomies and Faceted Search: Theory, Practice, and Experience* (The Information Retrieval Series). Berlin Heidelberg: Springer1-33 ..

Siddiqui, T., X. Ren, A. Parameswaran, & J. Han. 2016. Facetgist: Collective extraction of document facets in large technical corpora. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (pp. 871-880). Indianapolis, USA: 871-880

Stoica, E., M. A. Hearst and M. Richardson. 2007. Automating Creation of Hierarchical Faceted Metadata Structures. Paper presented at the HLT-NAACL. Rochester, NewYork: 244-251.

*UDC: Universal decimal classification*3 .rd ed. 2005. London: British Standards Institution.

Zheng, B., W. Zhang, & X. F. B. Feng. 2013. A survey of faceted search. *Journal of Web engineering*, 12 (1&2): 041-064.

Zolfaghar, Z., T. Mosavi Miangah, B. Rovshan, & A. R. Vakilifard. 2020. A Study on the Improved Techniques of Corpus-based Frequency Approaches in Automatic Term Extraction (ATE) (The Case Study: Basic Medicine Vocabulary). *Iranian Journal of Information processing and Management* 35 (4): 1039-1064.

عبدالحسین فرج‌پهلو

متولد سال ۱۳۳۰، دارای مدرک تحصیلی دکتری در رشتهٔ کتابداری و اطلاع رسانی از دانشگاه نیوساوت‌ویلز استرالیاست. ایشان هم‌اکنون استاد تمام گروه علم اطلاعات و دانش‌شناسی دانشگاه شهید چمران اهواز است.

مدیریت کتابخانه‌ها، مدیریت راهبردی در کتابخانه‌ها و مراکز اطلاع‌رسانی، مدیریت کیفیت، کاربرد فناوری اطلاعات و فناوری‌های نوین در کتابخانه‌ها و مراکز اطلاع‌رسانی از جمله علایق پژوهشی وی است.

فریده عصاره

متولد سال ۱۳۲۸، دارای مدرک تحصیلی دکتری کتابداری و اطلاع‌رسانی با گرایش علم‌سنجی و اطلاع‌سنجی از دانشگاه نیوسات‌ولز استرالیاست. ایشان هم‌اکنون استاد و مدیر قطب علمی مدیریت دانش و عضو گروه علم اطلاعات و دانش‌شناسی دانشگاه شهید چمران اهواز است.

تحلیل شبکه‌های اجتماعی هم‌نویسندگی، هم‌واژگانی و هم‌استنادی، دیداری‌سازی اطلاعات، بازیابی اطلاعات، داده‌کاوی، مطالعات بین رشته‌ای و هستی‌شناسی از جمله علایق پژوهشی وی است.

سید مصطفی فخراحمد

متولد سال ۱۳۵۹، دارای مدرک تحصیلی دکتری در رشتهٔ مهندسی رایانه، گرایش سیستم‌های نرم‌افزاری از شیراز است. ایشان هم‌اکنون دانشیار گروه مهندسی رایانهٔ دانشگاه شیراز است. پردازش زبان طبیعی، متن‌کاوی، مهندسی دانش و طراحی سیستم‌های اطلاعاتی تصمیم‌یار از جمله علایق پژوهشی ایشان است.

لیلا دهقانی

متولد ۱۳۵۷، ایشان هم‌اکنون استادیار گروه کتابداری در شاخهٔ پزشکی دانشگاه علوم پزشکی بوشهر است.

ذخیره و بازیابی اطلاعات، علم‌سنجی، داده‌کاوی و متن‌کاوی، رفتار اطلاعاتی و اخلاق نشر از جمله علایق پژوهشی ایشان است.