تحلیل توزیع و تمرکز کلیدواژه‌های پایان‌نامه‌ها و رساله‌ها: میزان تطابق با توصیفگرها، عنوان، و چکیده

اشکان خطیر
دانشجوی دکتری مهندسی فناوری اطلاعات
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

سهیل گنجه‌فر *
دکتری برق کنترل
استاد مهمان، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
استاد تمام دانشگاه بوعلی سینا دانشکده مهندسی، گروه برق

دریافت: ۹۵/۱۰/۲۹ پذیرش: ۹۶/۰۶/۲۵

چکیده: نمایه‌ها و چکیده‌های یک متن، خلاصه‌ای از متن را در اختیار خواننده قرار می‌دهند، لذا می‌توان از آن‌ها برای درک سریع و بازیابی سند استفاده کرد. ازآنجاکه بخش عمده‌ای از فعالیت‌های علمی فارسی در کشور ایران را پارساها تشکیل می‌دهند، در این پژوهش نمایه‌سازی پارساها از دو دیدگاه نویسنده پارسا و نمایه‌ساز حرفه‌ای موردبررسی قرار خواهد گرفت. سپس این نمایه‌ها با عنوان پارساها موردبررسی قرار می‌گیرد تا میزان انطباق با کلیدواژه‌های عنوانی به دست آید. از سوی دیگر با بررسی کلی مجموعه‌ای از نمایه‌ها و چکیده‌ها علاوه بر قابلیت بهبود در بازیابی اطلاعات برای محقق حوزه فعالیتی که بیشتر اسناد بر روی آن تمرکز کرده‌اند مشخص می‌شود.علاوه بر آن در این پژوهش وجود نمایه‌ها و توزیع آن‌ها در چکیده، بررسی می‌شوند. از توزیع کلیدواژه‌ها در چکیده می‌توان در استخراج خودکار کلیدواژه‌ها از چکیده پارساها در کارهای آتی استفاده شود. این پژوهش بر روی پارساهای موجود در پایگاه داده پژوهشگاه علوم و فناوری اطلاعات ایران که منبع گردآوری پارساهای فارسی است انجام شده است. روش پژوهش به این صورت است که بعد از گرداوری داده‌ها، پارساهایی که اطلاعات کافی ندارند پالایه شده و مابقی پارساها توسط برنامه‌ای که برای پردازش متن چکیده و نمایه‌های پارساها نوشته‌ایم مورد تحلیل قرار خواهند گرفت. سپس اطلاعات بدست آمده با استفاده از آمار توصیفی شرح داده خواهند شد. بررسی انجام‌شده در این پژوهش نشان داده است عموماً نمایه‌های انتخاب‌شده (بیش از ۶۰٪) توسط نویسنده و نمایه‌ساز حرفه‌ای از ۴۰٪ ابتدایی چکیده انتخاب‌شده‌اند. دیگر تحلیل‌های آماری این پژوهش نشان می‌دهند که میزان انطباق بین توصیفگرها و کلیدواژه‌ها ۸٪ است. این

فصلنامهٔ علمی پژوهشی
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا(چاپی) ۸۲۲۳-۲۲۵۱
شاپا(الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISA، SCOPUS وISC
http://jist.irandoc.ac.ir
دوره XX | شمارهٔ X | ص ص XX-XX
۱۳XX X

نوع مقاله: پژوهشی

به این مقاله به شکل زیر استناد کنید:
دورن متن:
(خطیر، اشکان ، زودآیند)
در فهرست منابع:
خطیر، اشکان. زودآیند. عنوان مقاله. تحلیل توزیع و تمرکز کلیدواژه‌های پایان‌نامه‌ها و رساله‌ها: میزان تطابق با توصیفگرها، عنوان، و چکیده *پژوهشنامه پردازش و مدیریت اطلاعات*.
http://Jipm.irandoc.ac.ir (دسترسی در روز/ماه/سال)

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

X xx ۱۳ | دورهٔ xx | شمارهٔ x

اختلاف نشان‌دهنده میزان تفاوت نظر زیاد بین نویسندگان پارساها و نمایه‌سازان است. با بهره‌گیری از این اختلاف و با تجمیع کلمات و غنی کردن کلیدواژه‌های سیستم بازیابی اطلاعات می‌توان در بهبود بازیابی اطلاعات نیز استفاده کرد.

کلیدواژه‌ها: نمایه‌سازی، کلیدواژه، توصیفگر، توزیع کلیدواژه، تمرکز فعالیت پژوهشی

*پدید آور رابط: دکتر سهیل گنجه‌فر Genjefar@irandoc.ac.ir ، S_ganjefar@basu.ac.ir

۱. مقدمه

امروزه به خاطر تولید حجم زیاد مستندات و بزرگ شدن پایگاه داده‌های نیمه ساخت‌یافته و غیر ساخت‌یافته، امکان بررسی تک‌تک این سندها برای پیدا کردن اطلاعات موردنیاز توسط کاربر و محقق امری بسیار دشوار، زمان‌بر و در عمل نشدنی است. حتی در محیط وب باآنکه موتورهای جستجو ابزارهای ارزشمندی را برای بازیابی اطلاعات در اختیار کاربر قرار می‌دهند اما در بسیاری از موارد ممکن است نتوانند نتایج مؤثری ارائه نمایند (Carpineto, Osiński, Romano, &Weiss, 2009).در بعضی موارد موتورهای جستجو افراد را به صفحاتی هدایت می‌کنند که ازنظر معنایی ارتباط زیادی با کلمات مورد جستجو ندارند. بنابراین برای دستیابی به نتایج مؤثر، روشی موردنیاز است تا کاربر نیاز نداشته باشد تا تمام صفحه وب، سند و یا بخش بزرگی از آن را موردمطالعه و بررسی قرار دهد و با صرف کمترین زمان بتواند اسنادی که ارتباط بیشتری را با موضوع و محتوای کاری خود دارد را پیدا کند. کلیدواژه کلمه و یا مجموعه کلماتی است که یک مفهوم عمیق و خلاصه‌ای از متن را ارائه می‌کند و نیز منجر به بازیابی اطلاعات خواهد شد(Hulth, 2003; Rose, Engel, Cramer, & Cowley, 2010). و در این مقاله منظور ما از توصیفگر، کلمه یا مجموعه کلماتی است که توسط نمایه‌ساز برای نشان دادن محتوای متن برای نمایه کردن آن متن انتخاب می‌شود. نمایه نیز کلمه یا کلماتی است که توسط نمایه‌ساز و یا نویسنده متن برای نمایه کردن و نشان دادن محتوای سند مورد انتخاب قرار می‌گیرند. بنابراین می‌توان با انتخاب مناسب کلیدواژه‌ها، و توصیفگرها سیستمی با بازیابی مناسب‌تری داشت. در این پژوهش نیز به بررسی و تجزیه‌وتحلیل آن‌ها خواهیم پرداخت.

انتخاب نمایه‌های مناسب علاوه بر اینکه در درک خواننده از متن کمک می‌کنند، کاربردهای فراوان دیگری نیز دارند. استخراج کلیدواژه‌ها در بعضی روش‌های خلاصه‌سازی متن به‌منظور استخراج ویژگی‌های متن و وزن دهی به کلمات، برای تعیین جملات پراهمیت

مورداستفاده قرارگرفته است (برای مثال نگاه کنید به (Chen, Han, & Chen, 2002; Gupta & Lehal, 2010; Kyoomarsi, Khosravi, Eslami, Dehkordy, & Tajoddin, 2008; Litvak & Last, (2008). از کاربردهای دیگر استخراج کلیدواژگان می‌توان به طبقه‌بندی خودکار متن (برای مثال (Zhang, 2008)) و نظر کاوی به‌عنوان استراتژی جستجو اشاره کرد (برای مثال (Khan, Baharudin, & Khan, 2009).

همان‌طور که ملاحظه کردیم، انتخاب کلیدواژه، کاربردی و حائز اهمیت است. بنابراین تحلیل و بررسی نمایه‌ها که از دید نویسنده و نمایه‌ساز برای یک سند انتخاب‌شده است ارزشمند است. برای اینکه یک سیستم بازیابی اطلاعات مناسبی داشته باشیم نیاز به درک درست و مناسبی از روابط نمایه‌ها و توزیع آن‌ها در عنوان و چکیده سند هستیم. این اطلاعات می‌تواند به طراحی سیستمی که به‌طور خودکار کلیدواژه‌ها را از متن انتخاب کند یاری رساند. از طرفی دیگر ازآنجاکه انتخاب نامناسب کلیدواژه‌ها مشکلاتی در بازیابی اطلاعات به وجود می‌آورد توصیه‌هایی مبنی بر عدم انتخاب کلیدواژه‌ها از عنوان نیز شده است تا در بازیابی اطلاعات بیشتر کمک کند. ازاین‌رو در این تحقیق به بررسی وجود یا عدم وجود نمایه‌ها در عنوان پارسا نیز پرداخته خواهد شد.

پژوهش‌هایی نیز در خصوص میزان انطباق سر عنوان‌های موضوعی با عناوین کتاب‌های منتشرشده نیز در کتابخانه ملی ایران انجام‌شده که نشان‌دهنده میزان انطباق ۹۹٪یی آن‌ها است (اباذری و پالیزوانی ۱۳۹۱). «اسکولتز» در پژوهش خود نشان داده است که میزان انطباق کلیدواژه‌های نویسنده به کلیدواژه‌های کنترل‌شده بسیار بیشتر از کلیدواژه‌هایی هستند که از عنوان استخراج‌شده‌اند (Schultz, Schultz, & Orr, 1965). «استریدر» نیز در (Strader, 2011) نشان داده است که کلیدواژه‌هایی که توسط نویسنده مشخص می‌شوند اطلاعات بسیار باارزش، اضافه بر نمایه‌ساز در اختیار قرار می‌دهند. در سال ۲۰۰۵ پژوهشی که توسط «انصاری» بر روی پایان‌نامه‌های پزشکی فارسی انجام شد، نشان می‌دهد که بیش از ۷۰٪ کلیدواژه‌ها عنوان و توصیفگرها با یکدیگر مطابقت دارند. (Ansari, 2005). در سال ۱۳۹۶ پژوهشی که توسط «قنواتی» بر روی مقاله‌های اریک و مندلی انجام شد، نشان می‌دهد که میزان شباهت بین کلیدواژه‌های نمایه‌ساز و نویسنده حدود ۴٪ است (قنواتی و همکاران، ۱۳۹۶).

در سال ۲۰۱۵ برای بررسی سر عنوان موضوعی پزشکی (MeSH) و تطابق آن با مجله‌های داروسازی که در مدلاین نمایه‌سازی شده‌اند پژوهشی انجام‌شده است. نتایج این پژوهش نشان داد ۴/۷ درصد مقاله‌هایی که بازیابی شده بودند از اصطلاحات مش استفاده نکرده‌اند و بطورکامل نمایه نشده بودند. پژوهش نشان داد که میانگین واژه‌های مش به ازای هر مقاله ۰/۹ بوده است. در

حدود ۵۲٪ از میان مقاله‌هایی که در ژورنال‌های داروسازی نمایه شده بودند نیز به‌طور کل از واژه‌های مش استفاده نکرده بودند (Minguet, Salgado, van den Boogerd, & Fernandez-Llimos, 2015). این بدان معناست که بیش از نیمی از مقاله‌های چاپ‌شده در ۱۰ ژورنال برتر حتی بدون استفاده از یک واژه مش نمایه شدند.

«مارگارت» (۲۰۱۱) در پژوهش خود به بررسی نمایه‌سازی محتوای آنلاین از سه دیدگاه گروه خوانندگان، نویسنده و نمایه‌سازان حرفه‌ای پرداخته است. میزان انطباق بین کلیدواژه نویسنده و برچسب‌های خوانندگان مقاله و کلیدواژه‌ها ۳۳ درصد، میزان انطباق بین توصیفگرها و برچسب‌ها ۱۶ درصد و میزان انطباق بین کلیدواژه‌ها و توصیفگرها ۱۹ درصد بوده است. «مارگارت» در پژوهش خود به مقایسه بین کلیدواژه‌ها، توصیفگرها و برچسب‌هایی که توسط کاربران داده‌شده‌اند پرداخته است و برچسب‌ها و نمایه‌های پرکاربرد را مشخص نموده و میزان آن‌ها را به ازای هر مقاله نشان داده است(Kipp, 2011). از طرفی در پژوهش دیگر خود نشان داده است که اگرچه بسیاری از برچسب‌ها با کلیدواژه‌ها و توصیفگرها شباهت دارند اما به‌طور کل با یکدیگر متفاوت می‌باشند (Kipp, 2006).

در پژوهش دیگری که توسط «بنی‌اقبالی» در ۲۰۱۱ بر روی پارسا‌های فارسی انجام‌شده است، واژه‌های عنوان و چکیده با توصیفگرهای تعیین‌شده در نمایه سازمان اسناد و کتابخانه ملی ایران با یکدیگر مقایسه شده است. نتایج این تحقیق نشان داد که میزان مطابقت واژه‌های عنوان با توصیفگرهای موجود در نمایه سازمان، ۴۷ درصد و میزان مطابقت واژه‌های چکیده با توصیفگرها ۵۳/۵٪ درصد بوده است (بنی‌اقبال و پیرهادی ۱۳۹۰)

در این پژوهش ابتدا رابطه بین کلیدواژه‌های نویسنده و نمایه‌ساز بررسی خواهد شد. سپس نمایه‌هایی که بیشترین تکرار را دارند مشخص می‌شوند و نشان داده می‌شود که چه میزان از نمایه‌ها در عنوان و چکیده پارساها قرار دارند. در ادامه پارساهایی که بیشترین نمایه را دارند مشخص خواهند شد. در انتها نیز توزیع کلیدواژه‌ها و توصیفگرها را در داخل چکیده موردبررسی قرار خواهند گرفت.

۲. هدف و سؤالات پژوهش

انتخاب کلیدواژه و نگارش چکیده برای یک متن علمی می‌تواند متخصصان و پژوهشگران را در بازیابی و درک بهتر متن یاری کند. تجربه نشان داده است که نظرات نویسنده سند و نمایه‌سازان در انتخاب نمایه برای یک سند متفاوت است. بنابراین می‌توان نتیجه گرفت درصورتی‌که از یکی از این نوع نمایه‌ها در سیستم بازیابی اطلاعات استفاده شود کارایی آن

سیستم کاهش می‌یابد. بنابراین مطالعه رابطه بین این دو می‌تواند در بازیابی اطلاعات بسیار مفید باشد. از طرفی دیگر در نمایه‌سازی خودکار که توسط رایانه صورت می‌گیرد، خصوصیت‌های ظاهری و آماری یک متن اولین ویژگی‌هایی هستند که توسط یک سیستم رایانه‌ای موردبررسی و تحلیل قرار می‌گیرند. هدف از انجام این پژوهش به دست آوردن، بررسی و تحلیل این اطلاعات آماری است تا علاوه بر اینکه معایب در انتخاب نمایه و نگارش چکیده کشف شود، اطلاعاتی نیز در خصوص نحوه انتخاب کلیدواژه‌ها ارائه شود که این اطلاعات می‌توانند در طراحی سیستمی برای استخراج خودکار کلیدواژه‌ها مورداستفاده قرار گیرند. مهم‌ترین سؤالات این پژوهش عبارت‌اند از:

- رابطه و الگوی کلیدواژه‌های انتخابی نویسنده و نمایه‌ساز چگونه است؟
- تمرکز تحقیقات بیشتر بر روی چه موضوعاتی است؟
- چه میزان کلیدواژه‌های انتخابی دانشجو در عنوان پارسا قرار دارند؟
- توزیع کلیدواژه‌ها در چکیده به چه صورت است؟

۳. روش تحقیق پژوهش

۳. ۱. نحوه انتخاب رساله، تعداد پارساها

برای انجام این تحقیق، از پارساهای فارسی بین سال‌های ۱۳۹۰ تا ۱۳۹۳ که در سامانه گنج ایرانداک ثبت‌شده‌اند، استفاده‌شده است. یکی از معیارهای انتخاب پارساها وجود کلیدواژه‌های نویسنده پارسا و نمایه‌ساز حرفه‌ای بوده است. ازآنجاکه رکوردهای رشته‌های مهندسی در بین این رکوردها بیشتر بودند از بین پارساهای بازیابی شده تنها پارساهای رشته‌ی مهندسی انتخاب‌شده است. تعداد این پارساها ۷۴۲ عدد می‌باشند. پارساهای انتخاب‌شده می‌بایست شامل مواردی همچون چکیده، کلیدواژه‌های نویسنده، و توصیفگرها باشند. ازآنجاکه پارساهای رشته‌های مهندسی در این مجموعه ازنظر تعداد با یکدیگر یکسان نبودند، پارساهای گرایش‌های مهندسی که تعداد بیشتری بودند انتخاب‌شده‌اند. پارساهای گرایش‌های مهندسی انتخاب‌شده شامل رشته‌های مهندسی برق، مهندسی مکانیک و مهندسی عمران می‌باشند.

۳. ۲. تحلیل داده‌ها

برای انجام پردازش و تحلیل، پارساهای مهندسی که در بین سال‌های ۱۳۹۰ تا ۱۳۹۳ در سامانه گنج به ثبت رسیده‌اند را در یک فایل اکسل ذخیره کردیم. هر رکورد از داده‌های ذخیره‌شده درنهایت حداقل شامل: عنوان، چکیده، کلیدواژه نویسنده، توصیفگر نمایه‌ساز،و فیلد تحصیلی است. برای انجام تحلیل‌های آماری از زبان برنامه‌نویسی C# استفاده‌شده است . مهم‌ترین

تحلیل‌های بکار گرفته‌شده در این بخش شامل موارد زیر می‌باشند:

•میزان اشتراک نمایه‌های نویسنده و نمایه‌ساز.

•تعداد نمایه‌های استفاده‌شده در هر رشته توسط نویسنده و نمایه‌ساز.

•نمایه‌های پرتکرار (و به‌تبع آن موضوعاتی که تمرکز بر روی آن‌ها است) و عناوین پارساهای دارای نمایه‌های پرتکرار.

•توزیع نمایه‌های استفاده‌شده در عنوان و چکیده.

٤. یافته‌ها

در این بخش یافته‌های پژوهش بر اساس ترتیب پرسش‌های پژوهش ارائه می‌شود. از بین پارساهایی که در سال‌های ۱۳۹۰ تا ۱۳۹۳ در پایگاه گنج ثبت‌شده‌اند، تنها ۱۹۸۴ پارسا هم دارای کلیدواژه و هم توصیفگر می‌باشند. از بین این ۱۹۸۴ پارسا ۷۴۲ پارسا مهندسی بوده و از بین آن‌ها ۵۲۷ پارسا در رشته‌های مهندسی برق، مهندسی مکانیک، و مهندسی عمران می‌باشند. جدول ۱ این اطلاعات را به‌صورت جزئی‌تری نشان می‌دهد:

جدول ۱) پایگاه داده موجود

	کل رشته‌ها	رشته‌های مهندسی	مهندسی برق	مهندسی مکانیک	مهندسی عمران	سه رشته
تعداد رساله	۱۹۸۴	۷۴۲	۲۰۴	۱۳۶	۱۸۷	۵۲۷
تعداد دانشگاه	۹۸	۶۷	۲۸	۳۳	۴۱	۵۰

درمجموع از ۵۲۷ پارسا که از سه رشته مهندسی عمران، مکانیک، و برق در جامعه آماری انتخاب‌شده‌اند، نتایجی به شرح زیر حاصل گردید. در سؤال اول این مطرح شده است رابطه و الگوی کلیدواژه‌های انتخابی نویسنده و نمایه‌ساز چگونه است؟ نتایج به‌دست‌آمده نشان داده است که به‌طور متوسط برای هر پارسا ۳/۲ کلیدواژه توسط نویسنده و ۷/۳ توصیفگر توسط نمایه‌ساز انتخاب‌شده است. آمار تعداد کلیدواژه‌ها و توصیفگرها در جدول ۲ آورده شده است:

جدول ۲) میزان کل نمایه‌هایی که توسط نویسنده و نمایه‌ساز مورداستفاده قرارگرفته‌اند

میزان اشتراک	کل کلمات	نمایه‌ساز	نویسنده	
۳۰۰ کلمه = ٪۸	۳۸۳۳	۲۵۶۰	۱۵۷۳	کلمات یکتا
-	۵۵۸۳	۳۸۶۴	۱۷۱۹	کل کلمات
-	۷,۲۷	۴,۸۶	۲,۹۸	میانگین کلمات یکتا در هر رساله
-	۱۰,۵۹	۷,۳۳	۳,۲۶	میانگین کل کلمات در هر رساله

دیگر نتایجی که در راستای این تحقیق به دست آمد نشان می‌دهد که از بین پنج پارساهایی که بیشترین نمایه‌ها را دارند سه پارسا متعلق به رشته عمران می‌باشند. از طرفی دیگر نیز از بین ۵ پارسایی که بیشترین توصیفگر را دارند نیز سه پارسا متعلق به رشته عمران و پارساهای عمران جز بیشترین کلیدواژه‌ها (که توسط نویسنده متن تخصیص داده‌شده‌اند) را به خود اختصاص داده‌اند. جداول مربوط به این نتایج در پیوست آمده است.

دیگر نتایج به‌دست‌آمده نشان می‌دهد که رشته عمران بیشترین اشتراک بین کلمات کلیدی و توصیفگرها را داشته‌اند. به صورتی که از بین ده پارسایی که بیشترین اشتراک بین کلیدواژه‌ها و توصیفگرها را دارند شش پارسای آن متعلق به رشته عمران است. این امر می‌تواند نشان‌دهنده هم‌فکری بین نویسندگان و نمایه‌سازان باشد. جدول مربوط به این موضوع در پیوست آمده است. بیشترین اشتراک بین نمایه‌ساز و نویسنده ٪۷۵ است.

در سؤال سوم مطرح‌شده است که تمرکز تحقیقات بیشتر بر روی چه موضوعاتی است؟ همان‌طور که می‌دانیم کلیدواژه‌ها نماینده یک سند می‌باشند. جدول ۳ پرتکرارترین نمایه‌ها را برای این پارساها نشان می‌دهد که از میان ۱۵۷۳ کلیدواژه نویسندگان و ۲۵۶۰ توصیفگر بیشترین کلماتی که توسط نویسنده، توصیفگر و به‌طورکلی مورداستفاده قرارگرفته‌اند به‌صورت زیر می‌باشند:

جدول ۳) پرتکرارترین نمایه‌ها

نمایه‌های پرتکرار	تکرار	-	توصیفگرهای پرتکرار	تکرار	-	کلیدواژه‌های پرتکرار	تکرار

شبکه عصبی	۴۰	-	شبکه عصبی	۲۷	-	بهینه‌سازی	۱۵
بهینه‌سازی	۳۶	-	روش اجزای محدود	۲۳	-	شبکه عصبی (شبکه عصبی مصنوعی)	۱۳
روش اجزای محدود	۲۵	-	بهینه‌سازی	۲۱	-	الگوریتم ژنتیک	۸
شبیه‌سازی	۲۰	-	شبیه‌سازی	۱۸	-	نانوسیال	۸
الگوریتم ژنتیک	۱۷	-	انتقال گرما	۱۷	-	انتقال حرارت	۵
انتقال گرما	۱۷	-	بتن مسلح	۱۵	-	شبیه‌سازی	۴

در بین این پارساها نویسندگان در ۱۰۲ پارسا تنها یک کلیدواژه و در ۶۲ پارسا تنها دو کلیدواژه به پارساهای خود اختصاص داده‌اند. نمایه‌ساز نیز برای هیچ پارسایی کمتر از۴ کلیدواژه اختصاص نداده است و به‌طور کل حدوداً همه پارساها بیش از ۶ نمایه را دارا می‌باشند. جدول ۴ تعداد پارساها را برای هر تعداد نمایه نشان می‌دهد.

جدول ٤) تعداد پارساها به ازای کلیدواژه‌ها و توصیفگرها

تعداد واژگان نمایه (کلیدواژه، توصیفگر یا کل نمایه‌ها) که به یک رساله تخصیص داده‌شده است.	تعداد پارس‌ها (نمایه شده توسط نویسنده) کلیدواژه	تعداد پارس‌ها (نمایه شده توسط نمایه‌ساز) توصیفگر	مجموعه نمایه‌ها (مجموع کلیدواژه‌ها و توصیفگرها)
۱	۱۰۲	۰	۰
۲	۶۲	۰	۰
۳	۶۵	۰	۰

۴	۲۲۳	۵	۰
۵	۵۳	۴۴	۱
۶	۱۶	۱۰۶	۱۶
۷	۱	۱۵۴	۳۹
۸	۵	۱۱۶	۷۷
۹	۰	۵۹	۹۲
۱۰	۰	۳۱	۱۰۲
۱۱	۰	۸	۸۲
۱۲	۰	۲	۴۹
۱۳	۰	۰	۳۵
۱۴	۰	۱	۴
۱۵	۰	۱	۲۴
بیش از ۱۵	۰	۰	۶

در جدول ۴، ۵ پارسا بیشترین تعداد کلیدواژگان را دارند و در ۱۰۲ پارسا، نویسنده تنها یک کلیدواژه به آن اختصاص داده است. از جدول ۴ می‌توان مشاهده نمود که ۶۵٪ از پارساها بین ۳ تا ۵ کلیدواژه دارند. در جدول ۷ عنوان‌های این پارساها آورده شده است. این جدول نشان می‌دهد که بیشترین کلیدواژه‌ها برای پارساهای رشته برق می‌باشند.

معیار دیگری که برای بازیابی بهتر اسناد در برخی انتشارات مانند الزویر ، پروکست و امرالد برای انتخاب کلیدواژگان پیشنهادشده است، انتخاب کلماتی خارج از عنوان برای نمایه است (EmeraldGroup). برای سنجش این پارامتر کلیدواژگان که توسط نویسنده و نمایه‌ساز انتخاب‌شده‌اند را با کلمات عنوان پارسا مقایسه شده‌اند.

سؤالی سوم پژوهش مطرح می‌کند که چه میزان کلیدواژگان انتخابی دانشجو در عنوان پارسا

قرار دارند؟ جدول ۵ نشان می‌دهد که از بین ۵۲۷ پارسا موردبررسی حدوداً کلیدواژه‌های ۵۳ پارسا (۱۰٪ پارساها) با عنوان انطباق کامل داشته‌اند و در ۲۳۲ پارسا، کلیدواژه‌ها آن‌ها از عنوان استخراج‌نشده است. از طرفی دیگر در ۱۳۸ پارسا هیچ توصیفگری از عنوان استخراج‌نشده است و هیچ یک از پارسا توصیفگرهای آن کلاً از عنوان استخراج‌نشده است، جدول ۶.

جدول ۵) میزان اشتراک کلیدواژه‌های نویسنده با عنوان

میزان اشتراک با عنوان	۰ ٪	بین ۰ تا ۲۵٪	بین ۲۵ تا ۵۰٪	بین ۵۰ تا ۷۵٪	بین ۷۵٪ تا ۱۰۰٪
تعداد پارساها	۲۳۲	۱۰۰	۱۱۲	۳۰	۵۳

جدول ٦) میزان اشتراک توصیفگرها با عنوان

میزان اشتراک با عنوان	۰ ٪	بین ۰ تا ۲۵٪	بین ۲۵ تا ۵۰٪	بین ۵۰ تا ۷۵٪	بین ۷۵٪ تا ۱۰۰٪
تعداد پارساها	۱۳۸	۲۵۳	۱۳۰	۶	۰

برای تحلیل دقیق‌تر میزان انطباق کلیدواژه‌ها و توصیفگرها با عنوان پارسا و چکیده پارسا موردبررسی قرارگرفته‌اند، که مشخص می‌کند نویسنده و نمایه‌ساز چه میزان از کلمات داخل چکیده و عنوان استفاده کرده‌اند. بررسی انجام‌شده نشان می‌دهد که در ۱۶۰ پارسا کل کلیدواژه‌ها از چکیده و عنوان استفاده‌شده است (حدود ۳۰٪ پارساها). از طرفی جدول ۷ نشان می‌دهد که در ۵۰٪ از پارساها بیشتر کلیدواژه‌ها را از داخل چکیده و عنوان استخراج‌شده‌اند. در جدول ۸ نشان داده‌شده است که میزان انطباق توصیفگرها هم با چکیده و هم عنوان ۱٪ است و ۶٪ از پارساها کلاً توصیفگرهایشان در چکیده و در عنوان نبوده است.

جدول ۷) میزان اشتراک کلیدواژه‌های نویسنده با عنوان و چکیده

میزان اشتراک	۰ ٪	بین ۰ تا ۲۵٪	بین ۲۵ تا ۵۰٪	بین ۵۰ تا ۷۵٪	بین ۷۵٪ تا ۱۰۰٪

تعداد پارسها	۷۱	۴۶	۱۴۲	۱۰۸	۱۶۰

جدول ۸) میزان اشتراک توصیفگرها با عنوان و چکیده

میزان اشتراک	۰٪	بین ۰ تا ۲۵٪	بین ۲۵ تا ۵۰٪	بین ۵۰ تا ۷۵٪	بین ۷۵٪ تا ۱۰۰٪
تعداد پارساها	۳۳	۱۵۴	۲۴۷	۸۷	۶

از طرف دیگر رابطه همبستگی بین تعداد کلیدواژه‌های انتخابی توسط دانشجو و طول چکیده و عنوان پارسا موردبررسی قرار گرفت. این بررسی نشان داد که هیچ رابطه‌ی معناداری بین طول چکیده و طول عنوان با تعداد کلیدواژه‌ها انتخاب‌شده توسط دانشجو وجود ندارد و همبستگی آن‌ها برابر صفر است. برای محاسبه این همبستگی ابتدا بردارهای کلیدواژه و طول عنوان را به برنامه متلب داده و همبستگی آن‌ها را محاسبه کرده که بسیار نزدیک به ۰ شد. در محاسبه همبستگی درصورتی‌که مقدار دو بردار برابر ۱ شود یعنی آن دو بردار با یکدیگر رابطه مستقیم و درصورتی‌که برابر ۱- شوند بدان معناست که رابطه معکوس دارند. درصورتی‌که همبستگی دو بردار برابر ۰ شود بدان معناست که هیچ رابطه‌ای بین آن دو بردار وجود ندارد. همبستگی بین بردار تعداد کلیدواژه‌ها و طول عنوان نیز برابر صفر شد.

در سؤال آخر پژوهش به دنبال این هستیم که توزیع کلیدواژه‌ها در چکیده به چه صورت است؟ برای بررسی توزیع نمایه‌ها در پارساها ابتدا نمایه‌هایی که در داخل چکیده قرار دارند جدا گردیده و سپس مکان فیزیکی آن‌ها را برحسب کلمه داخل چکیده پیداشده و به تعداد کل کلمات آن چکیده تقسیم و در عدد ۱۰۰ ضرب شده‌اند تا مشخص شود که آن کلمه در کدام بخش از چکیده قرار دارد. نتیجه این توزیع می‌تواند در طراحی یک سیستم نمایه‌سازی ماشینی برای به دست آوردن احتمال بیشتر وجود نمایه در مکان‌های مختلف چکیده مورداستفاده قرار بگیرد. از طرفی دیگر طبق استاندارد «انسی/میزو » چکیده شامل ۵ بخش است که به ترتیب عبارت‌اند از: هدف چکیده؛ روش‌شناسی؛ نتایج؛ جمع‌بندی و اطلاعات اضافی (National Information Standards Organization (NISO), 2010). بنابراین درصورتی‌که برای هر بخش تقریباً یک سهم مساوی در چکیده در نظر گرفته شود با بررسی توزیع کلیدواژه‌ها در چکیده می‌توان حدوداً به میزان اهمیت هر بخش پی برد. همان‌طور که در جدول ۹ ملاحظه می‌شود، اکثر

نمایه‌هایی که در داخل چکیده وجود دارد در ۴۰ درصد ابتدایی چکیده واقع‌شده‌اند.

جدول ۹) توزیع نمایه‌ها در چکیده

	درصد کلیدواژه‌ها	درصد توصیفگرها
در ۲۰٪ اول چکیده قرار دارند	٪۴۰	٪۴۵
در ۲۰٪ دوم چکیده قرار دارند	٪۲۴	٪۱۹
در ۲۰٪ سوم چکیده قرار دارند	٪۱۸	٪۱۸
در ۲۰٪ چهارم چکیده قرار دارند	٪۱۱	٪۱۰
در ۲۰٪ پنجم چکیده قرار دارند	٪۷	٪۷

۵. تجزیه‌وتحلیل یافته‌ها و جمع‌بندی

بیشتر پژوهش‌های پیشین در زمینه مقایسه کلیدواژه‌ها و توصیفگرها بر روی مقالات انجام‌شده است. در این پژوهش به بررسی میزان تشابه کلیدواژه‌ها و توصیفگرها پرداخته و نشان داده‌شده است که شباهت بین کلیدواژه‌های نویسنده و توصیفگرهای نمایه‌ساز کمتر از ۸٪ است. اختلاف این دو نشان دهنده اختلاف نظر زیاد در انتخاب کلیدواژه برای متن بین نویسنده و نمایه‌ساز است. بنابراین نمی‌توان به تنهایی از یکی در سیستم بازیابی اطلاعات موثر استفاده کرد. از طرفی پژوهشی که توسط «قواتی» برروی مقاله‌های پایگاه‌اطلاعاتی اریک و مندلی انجام شد نشان داده است که شباهت بین توصیفگرها و کلیدواژه‌ها ۴٪ است.

همان‌طور که در جدول ۲ مشاهده گردید، تفاوت کلیدواژه‌های یکتا و کل که توسط نویسنده انتخاب‌شده است بسیار اندک است و این بدان معناست که نویسنده‌های مقاله دایره لغات بسیار متفاوتی از یکدیگر دارند. می‌توان نتیجه گرفت که استفاده تنها از توصیف‌گرهای نمایه‌ساز نمی‌تواند کافی باشد. بنابراین استفاده هر یک به تنهایی این امر می‌تواند باعث عدم کارایی لازم در استفاده از کلیدواژه‌ها نویسنده در بازیابی اطلاعات شود.

از طرفی از آنجا که کلیدواژه‌ها در بازیابی اطلاعات نقش کلیدی دارند می‌توان با تجمیع کلمات و غنی کردن کلیدواژه‌ها سیستم بازیابی اطلاعات کارامدتری بدست آورد. از طرفی دیگر

نشان می‌دهد که ۶۰٪ رساله‌ها کمتر از ۲۵٪ کلیدواژه‌های آن‌ها از عنوان انتخاب شده‌اند. نتایج ما در جدول نشان می‌دهد که حدود ۵۰٪ کلیدواژه‌های نویسنده عینا از داخل چکیده استخراج شده‌اند که در پژوهش (Strader, 2011) نیز نتیجه مشابه ۵۴٪ی گرفته شده است.

در ادامه میزان اشتراک نمایه‌های استفاده‌شده توسط نویسنده و نمایه‌ساز با عنوان و چکیده موردبررسی قرار گرفت. بررسی انجام‌شده نشان می‌دهد که کلیدواژه‌های نویسنده در ۶۲٪ از پارساها کمتر از ۲۵٪ با عنوان استفاده کرده‌اند که این مقدار در مقایسه با انتخاب ۷۰٪ کلیدواژه‌ها از عنوان در پژوهش «انصاری» بسیار کمتر است. از طرفی دیگر ۵۱٪ از کلیدواژه‌های پارساهای بررسی‌شده بیش از ۷۵٪ با چکیده و عنوان شباهت دارند که در پژوهش (Strader, 2011) نیز نتیجه ۵۴٪ی گرفته شده است این نتایج نشان می‌دهند که نویسندگان عموماً اکثر کلیدواژه‌های خود را عیناً از چکیده انتخاب می‌کنند.

در ادامه نمایه‌های پرکاربرد در رشته‌های عمران، برق و مکانیک که در پارساهای فارسی مورداستفاده قرارگرفته‌اند ارائه‌شده است. نتایج نشان داد که در هر سه رشته عمران، مکانیک و برق از نمایه‌های شبیه‌سازی، بهینه‌سازی، شبکه‌های عصبی و الگوریتم ژنتیک بسیار استفاده‌شده است. ازآنجاکه کلیدواژگان نماینده یک سند است، استفاده زیاد از این لغات به‌عنوان کلمات می‌توان بیانگر این باشد که تمایل نویسندگان بیشتر بر روی چه موضوعی بوده است (Zhang, 2008; Rose, Engel, Cramer, & Cowley, 2010). بنابراین از جدول ۳ می‌توان نتیجه گرفت که تمایل و تمرکز پارساهای نوشته‌شده در سال‌های ۱۳۹۰ تا ۱۳۹۳ بیشتر بر روی بهینه‌سازی با استفاده از الگوریتم‌های تکاملی بوده است. از این سو می‌توان حدودی از محدود پژوهش‌ها را درک کرد و پژوهش‌ها با نیازهای صنعت همسوتر ساخت.

در انتها به بررسی وضعیت توزیع نمایه‌ها در داخل چکیده پرداخته و نشان داده شد ۶۴٪ نمایه‌هایی که در داخل چکیده وجود دارند در ۴۰٪ بخش نخست می‌باشند. این آمار نشان می‌دهد که نویسنده بیشتر مطالب خود را در هدف و روش‌شناسی بیان می‌کند و با استفاده از این ویژگی سیستم‌های استخراج خودکار کلیدواژه و یا خلاصه‌سازهای متن می‌توانند به کلمات کاندیدی که به‌عنوان کلیدواژه از این بخش‌ها استخراج می‌کنند وزن بیشتری قرار دهند.

فهرست منابع

اباذری، زهرا، سعیده پالیزوانی . ۱۳۹۱. بررسی میزان همخوانی و روزآمدی سر عنوان‌های موضوعی فارسی با عناوین کتاب‌های منتشرشده در پایگاه کتابخانه ملی ایران در سال‌های ۱۳۸۴-۱۳۸۸. *نظام‌ها و خدمات اطلاعاتی* ۲ (۱): ۴۲-

.۲۷

بنی اقبال، ناهید؛ فریبرز خسرو، ؛ صدیقه پیرهادی. ۱۳۹۰. مقایسه واژه‌های عنوان و چکیده پایان‌نامه‌ها با توصیفگرهای تعیین‌شده در نمایه سازمان اسناد و کتابخانه ملی ایران. *مطالعات ملی کتابداری و سازمان‌دهی اطلاعات* ۸۶: ۱۴۷-۱۳۴.

قنواتی، مریم؛ ، علی‌رضا نوروزی؛مریم ناخدا؛ اشکان خطیر ۱۳۹۶.(در دست چاپ) بررسی میزان تطابق زبان نمایه‌سازان، نویسندگان و برچسب‌گذاران در پایگاه اطلاعاتی اریک و مندلی. پژوهشنامه پردازش و مدیریت اطلاعات.

Ansari, Mariam. "Matching between Assigned Descriptors and Title Keywords in Medical Theses." Library Review 54, no. 7 (2005): 410-14.

Carpineto, Claudio, Stanislaw Osiński, Giovanni Romano, and Dawid Weiss. "A Survey of Web Clustering Engines." ACM Computing Surveys (CSUR) 41, no. 3 (2009): 17.

Chen, Fang, Kesong Han, and Guilin Chen. "An Approach to Sentence-Selection-Based Text Summarization." Paper presented at the TENCON'02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, 2002.

EmeraldGroup. "How To... Ensure Your Article Is Highly Downloaded: What You Can Do Prior to Submission." Emerald Group Publishing Limited, http://www.emeraldgrouppublishing.com/authors/guides/promote/optimize1.htm

Ghanavati Maryam, Alireza Noruzi, Maryam Nakhoda, Ashkan Khatir. In Press. Consistency between descriptors, author-supported keywords and tags in the ERIC and Mendeley databases. *Journal of Information Processing and Management*.

Gupta, Vishal, and Gurpreet Singh Lehal. "A Survey of Text Summarization Extractive Techniques." Journal of Emerging Technologies in Web Intelligence 2, no. 3 (2010): 258-68.

Hulth, Anette. "Improved Automatic Keyword Extraction Given More Linguistic Knowledge." Paper presented at the Proceedings of the 2003 conference on Empirical methods in natural language processing, 2003.

Khan, Khairullah, Baharum B Baharudin, and Aurangzeb Khan. "Mining Opinion from Text Documents: A Survey." Paper presented at the Digital Ecosystems and Technologies, 2009. DEST'09. 3rd IEEE International Conference on, 2009.

Kipp, Margaret EI. "Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator and Intermediary Keywords.(۲۰۰۶) ".

" .———Tagging of Biomedical Articles on Citeulike: A Comparison of User, Author and Professional Indexing." Knowledge Organization 38.(۲۰۱۱)

Kyoomarsi, Farshad, Hamid Khosravi, Esfandiar Eslami, Pooya Khosravyan Dehkordy, and Asghar Tajoddin. "Optimizing Text Summarization Based on Fuzzy Logic." Paper presented at the Seventh IEEE/ACIS International Conference on Computer and Information Science, 2008.

Litvak, Marina, and Mark Last. "Graph-Based Keyword Extraction for Single-Document Summarization." Paper presented at the Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, 2008.

Minguet, Fernando, Teresa M Salgado, Lucienne van den Boogerd, and Fernando Fernandez-Llimos. "Quality of Pharmacy-Specific Medical Subject Headings (Mesh) Assignment in Pharmacy Journals Indexed in Medline." Research in Social and Administrative Pharmacy 11, no. 5 (2015): 686-95.

National Information Standards Organization (NISO). "Ansi/Niso Z39.14-1997 (Guidelines for Abstracts)." National Information Standards Organization, 2010.

Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley. "Automatic Keyword Extraction from Individual Documents." Text Mining (2010): 1-20.

Schultz, Claire K, Wallace L Schultz, and Richard H Orr. "Comparative Indexing: Terms Supplied by Biomedical Authors and by Document Titles." American Documentation 16, no. 4 (1965): 299-312.

Strader, C Rockelle. "Author-Assigned Keywords Versus Library of Congress Subject Headings." Library resources & technical services 53, no. 4 (2011): 243-50.

Zhang, Chengzhi. "Automatic Keyword Extraction from Documents Using Conditional Random Fields." Journal of Computational Information Systems 4, no. 3 (2008): 1169-80.

پیوست

جدول ۱) بیشترین نمایه‌ها و عنوان رساله

عنوان رساله	رشته	تعداد نمایه‌ها
بررسی رفتار دینامیکی غیرخطی اتصال صلب، با مقطع کاهش‌یافته تیر(RBS) با روش اجزاء محدود	عمران	۱۹
تحلیل عددی بارگذاری دینامیکی بر روی یک سازه تحت محیط‌های واسط متفاوت	مکانیک	۱۹
شبیه‌سازی دوبعدی جریان در لوله‌ها	عمران	۱۸
آنالیز حساسیت شبکه GPS شمال تهران به فعالیت‌های احتمالی آتش‌فشان دماوند	عمران	۱۷
ارزیابی کفایت سیستم توزیع شامل واحدهای تولید پراکنده تجدید پذیر و خودروهای برقی	برق	۱۶

جدول ۲) پارساهایی که بیشترین توصیفگر ها را دارند

تعداد توصیفگر	رشته	عناوین پر توصیفگر نمایه‌ساز
۱۵	مکانیک	تحلیل عددی بارگذاری دینامیکی بر روی یک سازه تحت محیط‌های واسط متفاوت
۱۴	عمران	بررسی رفتار دینامیکی غیرخطی اتصال صلب، با مقطع کاهش‌یافته تیر(RBS) با روش اجزاء محدود
۱۲	برق	کنترل غیرخطی پدیده ی آشوب در دینامیک وضعیت ماهواره

۱۲	عمران	بررسی عملکرد لرزه‌ای قابهای با دهانه بلند و ارتفاع کم با عدم رعایت ضابطه تیر ضعیف-ستون قوی
۱۱	عمران	تعیین فرکانس ارتعاشی صفحات ترک‌دار با استفاده از روشهای بدون شبکه

جدول ۳) پارساهایی که بیشترین کلیدواژه ها را دارند

تعداد کلیدواژه	رشته	عناوین
۸	برق	ردیابی شی متحرک در شبکه‌های حسگر بیسیم با انرژی مصرفی کم
۸	برق	ارزیابی کفایت سیستم توزیع شامل واحدهای تولید پراکنده تجدید پذیر و خودروهای برقی
۸	عمران	آنالیز حساسیت شبکه GPS شمال تهران به فعالیت‌های احتمالی آتش‌فشان دماوند
۸	برق	طراحی و ساخت یک حسگر گاز بر مبنای ساختار فیبرنوری
۸	عمران	شبیه‌سازی دوبعدی جریان در لوله‌ها

جدول ٤) عناوین و تعداد کلمات کلیدی پارساهایی که کلمات کلیدی و توصیفگرهای آن‌ها بیشترین میزان اشتراک را با یکدیگر دارند

تعداد توصیفگر	تعداد کلیدواژه	تعداد کل کلمات	درصد اشتراک با کل نمایه‌ها	درصد کلمات مشترک با نویسنده	درصد کلمات مشترک با توصیفگرها	رشته	عنوان رساله
۶	۴	۷	٪۴۳	٪۷۵	٪۵۰	برق	بهبود کیفیت خدمات در سیستم‌های

							رادیوشناختی با استفاده از بهینه‌سازی برخی کمیت‌ها
۶	۴	۷	٪۴۳	٪۷۵	٪۵۰	عمران	تحلیل لرزه‌ای توربین‌های بادی با در نظرگرفتن اثرات اندرکنش خاک و سازه
۶	۴	۷	٪۴۳	٪۷۵	٪۵۰	عمران	شناسایی و پایش پدیده گرد و غبار از داده‌های سنجش از دوری با استفاده از عمق اپتیکی
۶	۴	۷	٪۴۳	٪۷۵	٪۵۰	عمران	بررسی خوردگی و نفوذ یون کلر در بتن خودتراکم حاوی زئولیت و متاکائولین
۵	۶	۸	٪۳۸	٪۵۰	٪۶۰	عمران	تاثیر خصوصیات هیدرولیکی مصالح سدهای غیرهمگن بر پایداری آن
۷	۴	۸	٪۳۸	٪۷۵	٪۴۳	مکانیک	تحلیل ویسکوالاستیک تنش تورق در اتصالات چسبی

۵	۳	۶	٪۳۴	٪۶۷	٪۴۰	برق	تخصیص منابع در شبکه‌های چندکاربره رله‌ای بی‌سیم رقابتی با رهیافت نظریه بازی
۸	۴	۹	٪۳۴	٪۷۵	٪۳۸	مکانیک	بررسی فرآیند ادغام قطرات معلق از طریق شبیه‌سازی عددی با روش‌های حجم سیال و شبکه‌ی بولتزمن
۸	۴	۹	٪۳۴	٪۷۵	٪۳۸	عمران	بررسی اثرات ساختگاه و تخمین خصوصیات جنبش نیرومند زمین با استفاده از داده‌های میکروترمور و مقایسه نتایج آن با مدلسازی پروفیل خاک (مطالعه موردی: ناحیه غربی شهر بابل)
۸	۴	۹	٪۳۴	٪۷۵	٪۳۸	عمران	بررسی روش آزمایشگاهی پاکسازی آلودگی نفتی درمحیط‌های آبی (آبهای زیرزمینی)

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

x شمارهٔ xx | دورهٔ xx | ۱۳xx x

The Analysis of the Distribution and Focus of Keywords in Theses and Dissertations: The Compliance with Descriptors, Title, and Abstract

Ashkan Khatir
Ph.D. Candidate in Information Technology Engineering; Iranian Research Institute for Information Science and Technology (IRANDOC).

Soheil Ganjefar*
Visiting Lecturer; Iranian Research Institute for Information Science and Technology (IRANDOC).
Professor of Electrical Engineering; Department of Electrical Engineering, Faculty of Engineering, Bu-Ali Sina University.

Ashkan Khatir: khatir@students.irandoc.ac.ir
Corresponding author Soheil Ganjefar: Genjefar@irandoc.ac.ir, S_ganjefar@basu.ac.ir

Abstract: Index terms provided by authors and professional indexers are used in traditional information retrieval schemes. However, abstracts ideally contain the core message of a document. This can potentially give us the opportunities to use the abstracts to automatically extract index terms. The purpose of this work is to be used as a base or as the first stage in the automatic keyword extraction as well as the high-level perception of where the ongoing research is headed Iranian Theses and Dissertations (TDs). To achieve the aforementioned objectives, we studied on more than 500 samples in different engineering research area from 50 different universities: 1) the correlation between the authors and professional indexers keywords. We observed only 8% similarity between these two indices. 2) We studied the correlation between the index terms and words in abstract and title. We found that that 40% of author keywords are extracted from first 20% of the abstract (This figure changes to 45%

for professional indexer) and 24% from the second 20% (19% from the next 20%) This finding can be further used to narrow down the input dimensions for the various machine learning schemes for automatic keyword extraction. 3) Using some classification schemes it can be perceived that the most of the ongoing research in Iran is headed toward neural network and optimization.

Keywords: Indexing, Descriptors, Keyword Distribution, Research Focus Area

"اشکان خطیر: متولد ۱۳۶۴ دانشجوی دکتری در رشته مهندسی فناوری اطلاعات در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک). تحلیل روند، متن‌کاوی و داده کاوی از جمله علایق پژوهشی وی است."

"سهیل گنجه‌فر: متولد ۱۳۴۹ دارای مدرک تحصیلی دکتری در رشته برق کنترل است. ایشان هم اکنون استاد تمام دانشگاه بوعلی سینا، گروه برق است. ازجمله علاقمندی‌های ایشان، یادگیری ماشین و متن‌کاوی می‌باشد."