مقایسه عملکرد اشتراک در مآخذ و اشتراک واژگان عناوین استنادها در خوشه‌بندی فازی پروانه‌های ثبت اختراع[1]

آناهیتا کرمانی * | کارشناسی ارشد کتابداری و اطلاع‌رسانی

نرگس نشاط [2] | دانشیار، سازمان اسناد و کتابخانه ملی ج.ا. ایران

عباس حری[3] | استاد، گروه کتابداری و اطلاع‌رسانی، دانشگاه تهران

دریافت: ۱۳۹۰/۰۹/۲۴ | پذیرش: ۱۳۹۱/۰۲/۱۳

فصلنامهٔ علمی پژوهشی
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا(چاپی) ۸۲۲۳-۲۲۵۱
شاپا(الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISA، SCOPUS و ISC
http://jipm.irandoc.ac.ir
دوره ۲۸ | شمارهٔ ۲ | ص ص ۴۱۱- ۴۳۲
زمستان ۱۳۹۱

نوع مقاله: پژوهشی

چکیده: هدف پژوهش حاضر، ارزیابی و مقایسه خصیصه استناد در قالب اشتراک در مآخذ و خصیصه واژگان عناوین استنادها در خوشه‌بندی پروانه‌های ثبت اختراع است. بدین منظور، این بررسی به‌روش تجربی بر روی مجموعه‌ای شامل پروانه‌های ثبت اختراع یو.اس. مربوط به رده ۹۷۷/۷۷۴ از رده‌بندی یو.اس. انجام شد. ابتدا پروانه‌های ثبت اختراع با استفاده از خوشه‌بندی فازی C میانگین خوشه‌بندی شد و خوشه‌بندی‌های حاصل با استفاده از معیارهای ارزیابی مانعیت و جامعیت گسترش‌یافته بی‌کیوبد مورد ارزیابی قرار گرفت. یافته‌ها نشان داد که خوشه‌بندی پروانه‌های ثبت اختراع با استفاده از اشتراک در مآخذ، عملکرد بهتری را نسبت به خوشه‌بندی با استفاده از واژگان عناوین استنادها دارد و ساختار خوشه‌ای در سطوح گسترده‌تری از جامعیت در خوشه‌بندی با استفاده از اشتراک در مآخذ برقرار است.

کلیدواژه‌ها: پروانه‌های ثبت اختراع، خوشه‌بندی، اشتراک در مآخذ، واژگان عناوین استنادها، خوشه‌بندی فازی C میانگین، ارزیابی خوشه‌بندی، مانعیت گسترش‌یافته بی‌کیوبد، جامعیت گسترش‌یافته بی‌کیوبد

۱. برگرفته از پایان‌نامه کارشناسی ارشد پدیدآور رابط با عنوان ارزیابی تأثیر استفاده از واژگان عناوین استنادی در مقایسه با استفاده از اشتراک استنادی در خوشه‌بندی پروانه‌های ثبت اختراع.

*kermanianahita@gmail.com
2. narges_neshat@yahoo.com
3. riwash@yahoo.com

۱. مقدمه

پروانه‌های ثبت اختراع یکی از منابع مهم اطلاعاتی در صنعت به‌شمار می‌رود که اطلاعات موجود در آنها می‌تواند به اهداف گوناگون، مورد استفاده و جستجو قرار گیرد. از جمله این اهداف می‌توان «تعیین قابلیت ثبت»[1] اختراعات مورد ادعا و یا «ردّ اعتبار»[2] اختراعات ثبت‌شده به‌واسطه یافتن اختراعی مشابه اشاره کرد (Graf and Azzopardi 2008; Bonino, Ciaramella, and Corno 2010; Hideo, Azzopardi, and Vanderbauwhede 2010). اهمیت بازیابی دقیق پروانه‌های ثبت اختراع از آنجا ناشی می‌شود که گاه فقط بازیابی یک مورد مرتبط می‌تواند منجر به ردّ اعتبار ادعا(ها)ی اختراع دیگری شود. به‌عبارت دیگر، در نظام‌های پژوهشی (ماهوی) ثبت اختراع، ثبت یک اثر اختراع منوط به دارا بودن شرایطی از جمله تازگی و نو بودن[3] و غیربدیهی بودن[4] آن است(United States Patent and Trademark Office (USPTO) 2010). از جمله فعالیت‌هایی که برای بررسی چنین شرایطی صورت می‌گیرد، جستجو در آثار پیشین[5] است. همچنین، ممکن است اختراعی به ثبت رسیده باشد، اما برای تأیید یا ردّ اعتبار آن نیاز به جستجو در آثار پیشین باشد.

خوشه‌بندی روشی است که برای گروه‌بندی مدارک مشابه مورد استفاده قرار می‌گیرد و شیوه‌ای مناسب در بازیابی (Kang et al. 2007) و ترسیم نقشه‌های علمی (Leydesdorff 1987) فراهم می‌کند. در این روش، مدارک در گروه‌هایی از پیش تعیین‌نشده به‌نام خوشه قرار می‌گیرند، به‌طوری‌که مدارک مشابه در کنار یکدیگر و مدارک و ماهیت‌های نامشابه دور از یکدیگر قرار گیرند. عوامل گوناگونی در خوشه‌بندی مدارک مؤثر است. یکی از این عوامل مربوط به نوع خصیصه‌ای[6] است که به‌واسطه آن یک مدرک مورد بازنمایی قرار می‌گیرد (Dhillon, Kogan, and Nicholas 2003). به‌همین دلیل، انتخاب درست خصیصه برای مدارک در نتایج خوشه‌بندی آنها تأثیرگذار خواهد بود.

استناد یکی از پرکاربردترین خصیصه‌هایی است که برای بازنمایی پروانه‌های ثبت اختراع مورد استفاده قرار می‌گیرد. اهمیت استنادها در پروانه‌های ثبت اختراع از آن جهت است که از یک طرف از سوی افراد متخصصی به‌نام بازرس تعیین می‌شوند و از سوی دیگر، به‌دلیل مرتبط بودن با ادعا(های) اختراع انتخاب می‌شوند. این استنادها در واقع، گزارش‌هایی هستند که اختراعی بودن یک اثر را مورد تأیید قرار می‌دهند (United States Patent and Trademark Office

1. Patentablity
2. invalidity
3. novelty
4. non-obviousness
5. prior art
6. attribute

(USPTO) 2010)[1]. از این رو، می‌توانند به‌عنوان خصیصه‌ای مناسب برای بازنمون پروانه‌های ثبت ثبت اختراع مورد استفاده قرار گیرند.

استفاده از استناد در بازنمایی پروانه‌های ثبت اختراع می‌تواند به‌شیوه‌های گوناگون صورت پذیرد. یکی از این شیوه‌ها، استفاده از «استناد» در خوشه‌بندی است که برمبنای «اشتراک در مآخذ» انجام می‌شود. یکی دیگر از راه‌ها، استفاده از «استناد» در خوشه‌بندی است که برمبنای «واژگان عناوین استنادها» صورت می‌گیرد. بدیهی است هر روزه تعدادی اختراع به‌ثبت می‌رسد و مورد استناد نیز قرار می‌گیرد. این امر سبب افزایش حجم استنادها و در نتیجه عاملی تأثیرگذار در خوشه‌بندی آنها خواهد بود. به‌همین دلیل، پیامدهای این امر سبب می‌گردد که شکل‌های مختلف پردازش و خوشه‌بندی برمبنای خصیصه استنادی مورد مداقه قرار گیرد تا بهترین راه که با جامعیت و مانعیت بالاتری همراه است، انتخاب شود. استفاده از استناد در بازنمایی پروانه‌های ثبت اختراع می‌تواند به‌شیوه‌های گوناگون صورت پذیرد. از این رو، پژوهش حاضر بر آن است تا با ایجاد یک خوشه‌بندی مبنا در پروانه‌های ثبت اختراع، به ارزیابی خوشه‌بندی از طریق واژگان عناوین استنادها و خوشه‌بندی از طریق اشتراک در مآخذ بپردازد و در این راستا، به پرسش‌های زیر پاسخ دهد:

۱. میزان جامعیت و مانعیت گسترش‌یافته بی‌کیوبد در خوشه‌بندی پروانه‌های ثبت اختراع با استفاده از اشتراک در مآخذ چقدر است؟
۲. میزان جامعیت و مانعیت گسترش‌یافته بی‌کیوبد در خوشه‌بندی پروانه‌های ثبت اختراع با استفاده از واژگان عناوین استنادها چقدر است؟
۳. کدام‌یک از دو نوع خوشه‌بندی ایجادشده از عملکرد بهتری (با توجه به معیارهای یادشده) برخوردار است؟

۲. برخی تعاریف

۲-۱. خوشه‌بندی

خوشه‌بندی مدارک عبارت است از تقسیم مجموعه‌ای از مدارک به‌تعدادی طبقه نامشخص به‌نام «خوشه». هدف از خوشه‌بندی، قرار دادن مدارک در خوشه‌ها به‌گونه‌ای است که اعضای هر خوشه از یک سو با یکدیگر دارای بیشترین شباهت و از سوی دیگر، با اعضای سایر خوشه‌ها دارای کمترین شباهت باشند (Tan, Steinbach, and Kumar 2006).

---

۱. البته باید توجه داشت که نظام‌های ثبت اختراع در مراکز گوناگون متفاوت است و توصیف یادشده از استناد در مراکزی صدق می کند که در آن ثبت اختراع مستلزم بررسی ماهوی اختراع است.

برای خوشه‌بندی پروانه‌های ثبت اختراع (مدارک) لازم است نوعی خصیصه برای پروانه‌ها تعیین شود تا به‌واسطه آن پروانه‌ها مورد بازنمون قرار گیرند (Jain and Dubes 1988 cited in Jain, Murty, and Flynn 1999). پس از آن، میزان شباهت (نزدیکی) مدارک به‌یکدیگر براساس معیار شباهت، محاسبه می‌شود و مدارک مشابه در کنار یکدیگر و مدارک نامشابه دور از هم قرار خواهند گرفت.

گاه ممکن است ارزش خصیصه‌ها در مدارک گوناگون متفاوت باشد. برای قائل شدن این تفاوت می‌توان به هر خصیصه در هر مدرک وزنی متفاوت اختصاص داد. یکی از روش‌های وزن‌دهی به خصیصه‌ها، روش tf-idf است (*tf–idf* 2011). در این روش، وزن هر خصیصه از ضرب «بسامد حضور یک خصیصه در یک مدرک» (tf) در «لگاریتم تعداد همه مدارک بخش بر تعداد مدارک حاوی هر خصیصه» (idf) به‌دست می‌آید.

استفاده از فضای برداری، یکی از فنون بازنمایی مدارک است. در این روش، مدارک با استفاده از خصیصه‌های وزن داده‌شده (به‌عنوان مؤلفه‌های بردار هر مدرک) مورد بازنمایی قرار می‌گیرند. معیار شباهت کوسینوسی یکی از متداول‌ترین معیارها برای سنجش شباهت میان مدارک است (Larsen and Aone 1999 cited in Huang 2008). بنابر این معیار، اگر d_k و d_j دو مدرک باشند شباهت میان آنها (کوسینوس زاویه میان مدارک آنها) از فرمول زیر به‌دست می‌آید:

$$\mathrm{sim}(\mathrm{d_j}, \mathrm{d_k}) = \frac{\vec{\mathrm{d_j}}.\vec{\mathrm{d_k}}}{|\vec{\mathrm{d_j}}||\vec{\mathrm{d_k}}|}$$

که در آن $\vec{d_j}.\vec{d_k}$ ضرب داخلی بردار دو مدرک در یکدیگر و $|\vec{d_j}|$ و $|\vec{d_k}|$ اندازه بردارهای $\vec{d_j}$ و $\vec{d_k}$ هستند. ازآنجاکه وزن هر خصیصه، مؤلفه‌های بردار هر مدرک را تشکیل می‌دهد، محاسبه شباهت بین مدارک را می‌توان به‌طریق زیر محاسبه کرد:

$$\mathrm{sim}(\mathrm{d_j}, \mathrm{d_k}) = \frac{\sum_{i=1}^{n} \mathrm{w_{i,j}} \mathrm{w_{i,k}}}{\sqrt{\sum_{i=1}^{n} \mathrm{w_{i,j}^2}} \sqrt{\sum_{i=1}^{n} \mathrm{w_{i,k}^2}}}$$

در این صورت، فاصله میان مدارک برابر است با (Huang 2008):

$$distance = 1 - similarity$$

بنابراین، هرچه میزان شباهت بیشتر باشد، فاصله میان دو مدرک کمتر خواهد بود.

۲-۱-۱. خوشه‌بندی فازی c میانگین[1]

خوشه‌بندی مدارک از طریق الگوریتم‌های گوناگون امکان‌پذیر است. طراحی برخی از این الگوریتم‌ها به‌گونه‌ای است که یک مدرک فقط به یک خوشه تعلق خواهد گرفت. در مقابل، در برخی الگوریتم‌ها امکان تعلق یک مدرک به بیش از یک خوشه امکان‌پذیر است. گونه نخست، خوشه‌بندی ناهمپوشان و گونه دوم، خوشه‌بندی همپوشان نامیده می‌شود. خوشه‌بندی فازی c میانگین (c به‌معنی خوشه) نوعی خوشه‌بندی همپوشان است. استفاده از این الگوریتم در گروه‌بندی پروانه‌های ثبت اختراع این امکان را فراهم می‌کند که پروانه‌هایی که هم‌زمان به بیش از یک موضوع تعلق دارند بتوانند به بیش از یک خوشه نیز تعلق داشته باشند. در نتیجه، این الگوریتم نسبت به خوشه‌بندی‌های ناهمپوشان، تناسب بیشتری با ماهیت‌های مورد خوشه‌بندی (پروانه‌های ثبت اختراع) خواهد داشت.

برای ایجاد خوشه‌بندی فازی c میانگین مراحل زیر دنبال می‌شود (Wedding 2009):

۱. انتخاب تعداد دلخواه k خوشه برای مجموعه‌ای شامل N مدرک[2]، به‌طوری‌که k<N؛
که در اینجا K برابر است با ۷۵ خوشه و N برابر با ۷۱۷ پروانه ثبت اختراع است.
۲. ایجاد یک نقطه مرکزی آغازین به‌ازای هریک از k خوشه؛
۳. محاسبه فاصله هریک از N مدرک تا هریک از k خوشه؛
۴. تعیین یک عضویت فازی یا کسری در هریک از k خوشه برای هریک از N مدرک؛
۵. یافتن نقطه مرکزی جدیدی برای هریک از k خوشه از طریق یافتن میانگین وزنی رکوردها؛ و
۶. تکرار مراحل ۳، ۴، و ۵ تا هنگامی که تغییری در عضویت خوشه‌ای به‌وجود نیاید (یا تا هنگام رسیدن به معیار همگرایی).

۲-۱-۱-۱. تعیین درجه عضویت مدارک در هریک از خوشه‌ها

درجه عضویت فازی، از محاسبه فاصله‌های مدارک از مرکز خوشه‌ها به‌دست می‌آید.

---

1. Fuzzy C-means clustering

۲. در توصیف الگوریتم، از واژه data استفاده شده است، زیرا کاربرد خوشه‌بندی فقط به خوشه‌بندی مدارک محدود نیست. اما، از آنجا که در اینجا خوشه‌بندی برای مدارک استفاده شده واژه مدرک مورد استفاده قرار گرفته است.

هنگامی که مدرکی به مرکز یک خوشه نزدیک‌تر از مرکز دیگر خوشه‌ها باشد، درجه عضویت مدرک در آن خوشه بیشتر و هنگامی که مدرک دورتر باشد، درجه عضویت کمتری خواهد داشت. مجموع مقادیر عضویت فازی [یک مدرک] نیز برابر با یک است (Wedding 2009).

درجه عضویت از طریق فرمول زیر محاسبه می‌شود:

$$u_k = \frac{1}{\sum_{i=1}^{j} \left(\frac{d_k}{d_i}\right)^p}$$

که در آن k یکی از j خوشه، d فاصله، و p توان است. توان (p) خود از فرمول زیر به‌دست می‌آید:

$$p = \frac{2}{(m-1)}$$

که در آن m یک توان[1] فازی است به‌طوری‌که $1<m<\infty$ است. مقدار m بر درجه فازی بودن خوشه‌ها تأثیرگذار است. هنگامی که $m \to 1$ (m نزدیک به ۱ باشد)، درجه فازی شدن خوشه‌ها کمتر و هنگامی که $m \to \infty$ (m به‌سمت بی‌نهایت میل کند)، خوشه‌ها به‌طور فزاینده‌ای فازی خواهند شد. برای مقادیر نزدیک به بی‌نهایت، مقدار عضویت‌های فازی یک مدرک در همه خوشه‌ها به برابری می‌گراید. به‌طور معمول، برای m مقدار ۲ در نظر گرفته می‌شود (Wedding 2009).

۲-۲. ارزیابی خوشه‌بندی‌های همپوشان: جامعیت و مانعیت گسترش‌یافته بی‌کیوبد[2]

معیارهای مورد استفاده در ارزیابی خوشه‌بندی‌های همپوشان با خوشه‌بندی‌های ناهمپوشان تفاوت دارد(Amigo et al. 2008). جامعیت و مانعیت گسترش‌یافته بی‌کیوبد از معیارهای ارزیابی خوشه‌بندی‌های همپوشان هستند که از گسترش جامعیت و مانعیت بی‌کیوبد[3] به‌دست آمده‌اند (Amigo et al. 2008). معیارهای بی‌کیوبد در حوزه‌هایی چون تعیین مدارک هم‌مرجع (در اینجا، هم‌خوشه[4])، بازیابی اطلاعات، و استخراج اطلاعات مورد استفاده قرار می‌گیرد (Bagga and Baldwin 1998).

---

1. exponent
2. extended BCubed precision and recall
۳. براساس ایمیل دریافت‌شده در ۶ فوریه ۲۰۱۲ از کوهان سوجای کارلس (Cohan Sujay Carlos)، بی‌کیوبد (B-Cubed) ، به‌معنی سه b، نام الگوریتم این معیارهاست که از اولین حرف نام ارائه‌دهندگان الگوریتم‌ها یعنی Biermann, Bagga, and Baldwin برگرفته شده است.
4. cross-document coreferencing

برای محاسبه جامعیت و مانعیت گسترش‌یافته بی‌کیوبد، ابتدا باید به محاسبه مانعیت و جامعیت چندگانه[1] پرداخت. مانعیت چندگانه[2] برابر است با مینیمم میان تعداد هم‌رخدادی دو مدرک در «خوشه‌بندی» و هم‌رخدادی آنها در«خوشه‌بندی مرجع» بخش بر تعداد هم‌رخدادی آنها در تعداد هم‌رخدادی در «خوشه‌بندی» همچنین جامعیت چندگانه[3] برابر است با مینیمم تعداد هم‌رخدادی دو مدرک در «خوشه‌بندی» و تعداد هم‌رخدادی آنها در «خوشه‌بندی مرجع» بخش بر تعداد هم‌رخدادی آنها در «خوشه‌بندی مرجع». این تعاریف به‌بیان ریاضی به صورت زیر تعریف می‌شوند (Amigo et al. 2008):

$$Multiplicity\ Precision(e,e') = \frac{Min(|C(e) \cap C(e')|, |L(e) \cap L(e')|)}{|C(e) \cap C(e')|}$$

$$Multiplicity\ Recall(e,e') = \frac{Min(|C(e) \cap C(e')|, |L(e) \cap L(e')|)}{|L(e) \cap L(e')|}$$

که در آن e و e′ نشان‌دهنده دو مدرک است. L (e) مجموعه خوشه‌های مرجع و C (e) مجموعه خوشه‌های مربوط به e است. |C(e)∩C(e′)| تعداد هم‌رخدادی e و e′ در خوشه‌بندی و |L(e)∩L(e′)| تعداد هم‌رخدادی e و e′ در خوشه‌بندی مرجع است. مقدار مانعیت گسترش‌یافته بی‌کیوبد برای یک مدرک برابر است با میانگین مانعیت چندگانه آن مدرک نسبت به مدارک با خوشه مرجع مشترک و مقدار نهایی مانعیت گسترش‌یافته بی‌کیوبد برابر با میانگین مانعیت گسترش‌یافته بی‌کیوبد برای هر مدرک است. به‌همین ترتیب، مقدار جامعیت گسترش‌یافته بی‌کیوبد نیز از فرایندی مشابه محاسبه می‌شود. توضیحات یادشده به‌زبان ریاضی در ادامه آمده است (Amigo et al. 2008):

$$Precision\ BCubed = Avg_e\left[Avg_{e',C(e)\cap C(e')\neq\emptyset}[Multiplicity\ precision(e,e')]\right]$$

$$Recall\ BCubed = Avg_e\left[Avg_{e',L(e)\cap L(e')\neq\emptyset}[Multiplicity\ recall(e,e')]\right]$$

1. multiplicity precision and recall
2. multiplicity precision
3. multiplicity recall

۳. پیشینه پژوهش

در میان آثار فارسی، اثری که در آن خوشه‌بندی پروانه‌های ثبت اختراع را با استفاده از عناصر استنادی مورد ارزیابی قرار داده باشد یافت نشد؛ با این حال، می‌توان به اثر داروغه (۱۳۸۰) اشاره کرد که در پایان‌نامه خود به بررسی میزان همپوشانی کلیدواژه‌های عنوان و عناوین مآخذ با توصیفگرهای نمایه‌سازی در پایان‌نامه‌های دکترای تخصصی روان‌پزشکی، زنان و زایمان، و قلب و عروق دانشگاه علوم پزشکی ایران پرداخته است. یافته‌های این پژوهش نشان داد که میزان همپوشانی سرجمع کلیدواژه‌های عنوان و عناوین مآخذ با توصیفگرهای نمایه‌سازی در هر رشته روان‌پزشکی، قلب و عروق، و زنان و زایمان بیش از میزان همپوشانی کلیدواژه‌های عنوان با توصیفگرهای نمایه‌سازی است.

کسلر طی مقاله‌ای مطالعه‌ای مقدماتی برای آزمون فرضیه اشتراک در مآخذ انجام داد. این مطالعه بر روی ۴۰ مدرک منتشرشده در ژوئن ۱۹۵۸ در *مجموعه مقالات آی.آر.ایی*.[1] با موضوع ترانزیستورها صورت گرفت. ارزیابی گروه‌های به‌وجودآمده از طریق بررسی موردی برخی مقالات و مقایسه موضوعی آنها با دیگر مقالات موجود در آن گروه، از سوی نویسنده صورت گرفته است. او در نهایت، نتیجه گرفت که می‌توان بدین شیوه به ایجاد گروه‌های کوچک‌تری از مقالات پرداخت به‌گونه‌ای که هر گروه زیرمجموعه درستی از گروه بزرگ‌تر باشد (Kessler 1963).

سالتون استفاده از استنادهای مآخذ در مقایسه با استفاده از واژگان در بازیابی مدارک مرتبط را مورد مقایسه قرار داده است. این پژوهش بر روی ۶۲ مدرک در حوزه زبان‌شناسی و ترجمه ماشینی صورت گرفته است. شواهد پژوهش به‌روشنی نشان نداده است که استنادها در نظام‌های بازیابی خودکار مدارک به‌طور معنی‌داری نقش مؤثری ایفا می‌کنند (Salton 1963).

سالتون عملکرد بازیابی را هنگام اضافه کردن اطلاعات استنادی به بردارهای تشکیل‌شده از واژه‌ها و مفاهیم و نیز استفاده از اطلاعات استنادی به‌تنهایی در ایجاد بردارها، نسبت به حالتی که از استناد استفاده نمی‌شود (استفاده از واژگان و مفاهیم)، مورد مقایسه قرار داده است. این پژوهش، بر روی چکیده ۲۰۰ مقاله در حوزه آئرودینامیک به‌همراه ۴۲ پرسش اخذشده از پژوهشگران حوزه آئرودینامیک صورت گرفته است. نتایج نشان داده است که اضافه کردن مفاهیم استنادی به توصیفگرهای موضوعی موجب افزایش میزان مانعیت تا بیش از ۱۰ درصد در آستانه‌ای از جامعیت خواهد شد. همچنین، خصیصه‌های استنادی عملکرد بسیار بهتری در سطوح پایین جامعیت (مانعیت‌محوری) دارند (Salton 1971).

1. Proceedings of the IRE

شاو به ارزیابی حضور ساختار خوشه‌ای به‌عنوان تابعی از جامعیت[1] در مدارکی با موضوع فیبروز کیستی[2] پرداخته است. این ارزیابی برای چهار نوع بازنمایی موضوعی و دو نوع بازنمایی استنادی صورت گرفته است. نتایج نشان داد که برای تمامی انواع بازنمایی‌های مورد پژوهش، با کاهش جامعیت در بازنمایی، حضور ساختار خوشه‌ای کم خواهد شد. باوجود اینکه بسیاری از مدارک فاقد ارجاع یا استناد بوده‌اند، در بازنمایی استنادی حضور ساختار خوشه‌ای را در طیف گسترده‌تری از سطوح مختلف جامعیت نسبت به بازنمایی موضوعی نشان داد. هر دو نوع بازنمایی استنادی نشان‌دهنده شواهدی مبنی بر وجود ساختار خوشه‌ای در کمترین حالت جامعیت بود. ساختارهای اعمال‌شده در این مجموعه با استفاده از نمایه‌های موضوعی و استنادی، شرایط لازم را برای ایجاد خوشه‌بندی معنی‌دار فراهم کرده‌اند (Shaw 1990).

شاو میزان اثربخشی بازیابی مبتنی بر خوشه‌بندی را برای پنج نوع بازنمایی، مبتنی بر بازنمایی موضوعی و بازنمایی استنادی مورد بررسی قرار داده است. این آزمایش روی مجموعه‌ای شامل ۱۲۳۹ مدرک با موضوع فیبروز کیستی صورت گرفت که با اصطلاح Cystic Fibrosis در کتابخانه ملی پزشکی (مدلاین) نمایه شده بودند. ترکیبات مورد بررسی شامل استفاده از سرعنوان موضوعی و منابع مورد استناد، سرعنوان موضوعی و منابع استنادکننده، سرعنوان موضوعی پزشکی و منابع مورد استناد، سرعنوان موضوعی پزشکی و منابع استنادکننده، و منابع مورد استناد و منابع استنادکننده بوده است. نتایج این پژوهش نشان داد که مقدار بهینه عملکرد تمامی ترکیب‌ها، قابل مقایسه و بیش از مقدار بهینه عملکرد در بازنمایی به‌صورت غیرترکیبی[3] است. عملکرد بهینه، به‌طور ثابت، در سطوح پایینی از جامعیت‌محوری[4] رخ داده است. در میان بازنمایی‌های ترکیبی جامعیت‌محور[5]، مواردی که بازنمایی موضوعی را مورد استفاده قرار داده بوده‌اند، سطح عملکرد پایین‌تری داشته‌اند، نتایج بازیابی حاصل از ساختار تصادفی، قابل مقایسه با نتایج مشاهده‌شده بوده است. در میان بازنمایی‌های ترکیبی جامعیت‌محور، مواردی که منابع استناد کننده و استنادشونده را مورد استفاده قرار داده‌اند، نسبت به بازنمایی‌های حاصل از خصیصه‌های موضوعی، به‌طور قابل ملاحظه‌ای از اثربخشی بیشتری برخوردار بوده‌اند (Shaw 1991).

لای و وو در مقاله خود با عنوان «استفاده از رویکرد هم استنادی برای ایجاد نظامی نوین در رده‌بندی پروانه‌های ثبت اختراع» استفاده از رویکرد هم‌استنادی را به‌منظور رده‌بندی پروانه های ثبت اختراع پیشنهاد کرده‌اند. نتایج ارزیابی این مدل روی مجموعه‌ای آزمایشی شامل

1. Exhaustivity
2. cystic fibrosis (CF)
3. constituent representation
4. exhaustivity
5. exhaustive composite representation

پروانه‌های مرتبط با نانوتکنولوژی حاکی از مفید بودن اطلاعات استنادی در سازماندهی پروانه‌های ثبت اختراع بوده است (Lai and Wu 2005).

لی و همکاران در مقاله‌ای با عنوان «رده‌بندی خودکار پروانه‌های ثبت اختراع با استفاده از شبکه استنادی: مطالعه‌ای تجربی در فناوری نانو» استفاده از اطلاعات استنادی پروانه‌های ثبت اختراع به‌ویژه اطلاعات ساختار شبکه استنادی را در اختصاص خودکار رده‌ها به پروانه‌های ثبت اختراع مورد بررسی قرار داده‌اند. رویکرد مورد استفاده در پژوهش آنها مبتنی بر کرنل[1] بوده و برای اخذ اطلاعات محتوایی و استنادی از تعریف توابع کرنل با استفاده از مجموعه‌ای از مدارک حاوی پروانه‌های ثبت اختراع در حوزه نانوتکنولوژی صورت گرفته است. نتایج این پژوهش حاکی از آن است که استفاده از ساختارهای شبکه استنادی به‌طور معنی‌داری عملکرد بهتری نسبت به حالتی ایجاد می‌کند که در آن از استناد استفاده نشده است (Li et al. 2007).

فوجی در مقاله‌ای با عنوان «تلفیق اطلاعات استنادی و محتوایی در بازیابی ثبت اختراع ان.تی.سی.آی.آر.-۶» به ارزیابی بازیابی اطلاعات پروانه‌های ثبت اختراع با هدف ردّ اعتبار پروانه‌های ثبت اختراعات پرداخته است. مدل پیشنهادی او شامل تلفیق اطلاعات استنادی و متنی برای بازیابی پروانه‌های ثبت اختراع بوده است. در این مدل، ابتدا از بازیابی مبتنی بر متن و پس از آن، از اطلاعات استنادی استفاده است. در بازیابی مبتنی بر متن، از متن ادعای(های) موجود در پروانه‌های ثبت اختراع استفاده شده است. در نمره‌دهی استنادی مدارک دو شیوه در پیش گرفته شده است: حالتی که نمره استنادی در کل مدارک در نظر گرفته می‌شود و حالتی که نمره استنادی در N مدرک برتر حاصل از بازیابی مبتنی بر متن محاسبه می‌شود. در این مدل، به‌منظور نمره‌دهی ثبت اختراعات از نظر اطلاعات استنادی از روش پیج‌رنک[2] استفاده گردیده است. نتایج این پژوهش تجربی بر روی مجموعه آزمایشی پروانه‌های ثبت اختراع ان.تی.سی.آی. ر.-۶ نشان داده است که تلفیق اطلاعات استنادی در حالت اول (محاسبه نمره استنادی بدون در نظر گرفتن N مدرک برتر) اثربخش‌تر از حالت دوم عمل کرده است. اما به‌طور کلی، تلفیق اطلاعات استنادی و متنی در بازیابی اثربخش بوده است (Fujii 2007).

تیوانا و هورویتز در «فایندسایت - یافتن خودکار پروانه‌های ثبت اختراع پیشین»، به ارائه الگوریتمی پرداخته‌اند که در آن از استنادها برای یافتن پروانه‌های ثبت اختراع استفاده شده است. این الگوریتم بر روی نتایج بازیابی پیاده‌سازی می‌شود، به این معنی که ابتدا واژه‌ای جستجو و پس از آن الگوریتم بر روی مجموعه مدارک بازیابی‌شده، پردازش‌های بعدی را

1. Kernel 2. PageRank

انجام می‌دهد. ابتدا از طریق جستجوی کلیدواژه‌ای، مجموعه‌ای از پروانه‌های ثبت اختراع مشخص می‌شود. سپس، پروانه‌های ثبت اختراع مهم از طریق استنادهای پیش‌آیند و پس‌آیند مشخص می‌گردد. استفاده از استناد در الگوریتم پیشنهادی سبب بازیابی تعداد زیادی از منابع نامرتبط در کنار منابع مرتبط بوده است (Tiwana and Horowitz 2009).

زو و کرافت در مقاله خود با عنوان «تبدیل پروانه‌های ثبت اختراع به‌عبارت‌های جستجو برای آثار پیشین» به شناسایی بهترین ناحیه برای انتخاب خودکار عبارت جستجو پرداخته‌اند. نتایج پژوهش بر روی مجموعه حاوی پروانه‌های ثبت اختراع اداره یو.اس.پی.تی.اُ[1] حاکی از آن است که از میان نواحی عنوان، چکیده، خلاصه آثار پیشین[2]، توصیف اشکال، شرح جزئیات، ادعاها، جستجوی آزاد و ادعاهای مقدم ، ناحیه خلاصه آثار پیشین دارای بهترین تأثیر در بازیابی آثار پیشین بوده است. همچنین، در میان الگوریتم‌های مورد استفاده برای وزن‌دهی، الگوریتم tf دارای بهترین عملکرد و پس از آن، الگوریتم‌های tf-idf و bool قرار داشته‌اند (Xue and Croft 2009).

جستجو در آثار پیشین حاکی از آن است که موارد اندکی به بررسی نقش استناد در سازماندهی پروانه‌های ثبت اختراع پرداخته‌اند و توجه آنها نیز بیشتر بر رده‌بندی خودکار یا بازیابی پروانه‌های ثبت اختراع بوده است. با توجه به آنکه استناددهی در پروانه‌های ثبت اختراع از سوی متخصصان صورت می‌گیرد (United States Patent and Trademark Office (USPTO 2010)، احتمال وجود ارتباط میان استنادها و اختراع ادعاشده بیشتر خواهد بود و در نتیجه به‌نظر می‌رسد که خوشه‌بندی پروانه‌های ثبت اختراع برمبنای استناد نتایج مطلوبی دربرداشته باشد. به‌همین منظور، پژوهش حاضر بر آن است تا به ارزیابی خوشه‌بندی برمبنای «اشتراک در مآخذ» و «واژگان عناوین استنادها» بپردازد.

٤. روش پژوهش و ابزارهای گردآوری اطلاعات

در پژوهش حاضر که از نوع تجربی است، پس از استخراج و ذخیره اطلاعات در نرم‌افزار اکسل، ابزارهای زیر برای مراحل مختلف مورد استفاده قرار گرفت:

نرم‌افزار کد منبع باز رپیدماینر[3] در وزن‌دهی به واژگان و تشکیل ماتریس‌های واژه- مدرک؛ برنامه[4] خوشه‌بندی فازی[5] نوشته‌شده به زبان برنامه‌نویسی پرل[6] برای ایجاد خوشه‌بندی فازی c

1. USPTO 2. background summary 3. http://rapid-i.com/content/view/181/190/
4. http://cpan.uwinnipeg.ca/htdocs/Algorithm-FuzzyCmeans/Algorithm/FuzzyCmeans.html
5. fuzzy c-means clustering 6. Perl

میانگین و برنامه[1] ارزیابی‌کننده خوشه‌بندی براساس جامعیت و مانعیت گسترش‌یافته بی‌کیوبد که در کارگروه وپس[2] معرفی و ارائه گردیده است. علاوه بر نرم‌افزارها و الگوریتم‌های پیش‌گفته، در تهیه نمودارها نرم‌افزار اکسل و اس.پی.اس.اس.[3] مورد استفاده قرار گرفتند.

جامعه آماری شامل ۷۱۷ پروانه ثبت اختراع یو.اس.[4] مورد استناد در پروانه‌های ثبت اختراع یو.اس. متعلق به رده ۹۷۷/۷۷۴ (از رده‌بندی یو.اس.[5]) بوده است. در رده‌بندی پروانه‌های ثبت اختراع یو.اس. رده ۹۷۷ رده‌ای اصلی و مربوط به موضوع نانوتکنولوژی است. ۹۷۷/۷۷۴ رده فرعی و زیرمجموعه‌ای[6] از رده فرعی ۹۷۷/۷۷۳ است. رده ۹۷۷/۷۷۳ نیز زیرمجموعه‌ای از رده ۹۷۷/۷۰۰ است. رده ۹۷۷/۷۰۰ مربوط به نانوساختارها و رده ۹۷۷/۷۷۳ شامل نانوذرات است (*Class definition for class 977* 2011).

٤-١. نحوه ایجاد مجموعه مورد آزمایش[7] با استفاده از تعیین خوشه‌بندی مرجع

ارزیابی خوشه‌بندی به‌طور معمول، با استفاده از مجموعه‌های آزمایشی صورت می‌گیرد. مجموعه آزمایشی، دربرگیرنده مدارکی است که نسبت به آنها قضاوت‌های ربط صورت گرفته است. در این مجموعه‌ها، موضوعاتی تعریف شده است که تعدادی از مدارک با آنها مرتبط است.

خوشه‌بندی مرجع، گونه‌ای خوشه‌بندی است که در آن مدارک مرتبط به‌درستی در خوشه‌ها (و به‌عبارت دیگر، در کنار یکدیگر) قرار گرفته‌اند. بنابراین، پس از تطابق خوشه‌بندی پیشنهادی و خوشه‌بندی مرجع می‌توان به ارزیابی خوشه‌بندی پیشنهادی پرداخت.

یکی از روش‌های ایجاد این مجموعه‌ها برای پروانه‌های ثبت اختراع، استفاده از روابط استنادی است. در این روش، هر پروانه ثبت اختراع مانند موضوعی است که استنادهای موجود در آن، مدارک مرتبط را مشخص می‌کند (*Generation of a test collection based on citations* [n.d.]). بنابراین، هر پروانه مانند خوشه‌ای حاوی مجموعه‌ای از مدارک مرتبط است. در این صورت، مجموعه‌ای از پروانه‌های ثبت اختراع خود می‌تواند به‌عنوان خوشه‌بندی مرجع در نظر گرفته شود.

در پژوهش حاضر، پروانه‌های ثبت اختراع رده ۹۷۷/۷۷۴ به‌عنوان خوشه‌بندی مرجع و استنادهای صورت‌گرفته در آنها به‌عنوان مجموعه مورد آزمایش در نظر گرفته شده است. همچنین، ازآنجاکه خوشه‌بندی پروانه‌های ثبت اختراع یو.اس. مورد نظر بوده است از میان منابع مورد استناد، فقط پروانه‌های ثبت اختراع یو.اس. در مجموعه قرار گرفته است.

---

1. www.evalita.it/sites/evalita.fbk.eu/files/doc2011/weps2007_scorer_1.1.tar.gz
2. WePS
3. SPSS
4. US patent
5. US patent classification
6. indented
7. test collection

برای ایجاد این مجموعه ابتدا تعداد ۱۰۴ پروانه ثبت اختراع مربوط به رده ۹۷۷/۷۷۴ به‌عنوان خوشه‌بندی مرجع برگزیده شد که برای به‌دست آوردن آنها عبارت ccl/977/774 در جستجوی پیشرفته پایگاه مربوط به پروانه‌های ثبت اختراع اداره یو.اس.پی.تی.اُ. به نشانی http://patft.uspto.gov/netahtml/PTO/search-adv.htm مورد جستجو قرار گرفت. از میان ۱۰۴ خوشه (پروانه)، تعداد ۱۸ خوشه به‌دلیل یکسان بودن با پروانه‌های مورد استناد و تعداد ۱۱ خوشه نیز به‌دلیل عدم استناد به پروانه ثبت اختراع یو.اس. از خوشه‌بندی مرجع حذف شد. در نهایت، تعداد ۷۵ خوشه به‌دست آمد. تعداد پروانه‌های ثبت اختراع یو.اس. مورد استناد در این پروانه‌ها برابر با ۷۱۷ پروانه ثبت اختراع بوده است که در این پژوهش، به‌عنوان مجموعه آزمایشی مورد استفاده قرار گرفت.

٤-٢. استخراج خصیصه‌ها

پس از مشخص شدن مجموعه آزمایشی، اقدام به استخراج «شماره‌ها» و «عناوین پروانه های ثبت اختراع مورد استناد» در هر پروانه ثبت اختراع (با استفاده از پایگاه‌های اطلاعاتی یو.اس.پی.تی.اُ.، گوگل پتنتس[1]، فری پتنتس‌آنلاین[2]) و درج آنها در قالب جداولی در نرم‌افزار اکسل صورت گرفته است. در نهایت، دو جدول تشکیل شد: یک جدول برای ایجاد اشتراک در مآخذ و جدول دیگر برای استفاده از واژگان عناوین استنادها.

٤-٣. ایجاد ماتریس خصیصه - پروانه ثبت اختراع

برای ایجاد ماتریس«استناد- پروانه ثبت اختراع» جدول مربوط به اشتراک در مآخذ به نرم‌افزار رپیدماینر وارد شد و وزن‌دهی استنادها براساس روش *tf-idf* (*tf–idf* 2011) صورت گرفت.

برای ایجاد ماتریس واژگان عناوین استنادی– پروانه ثبت اختراع نیز از جدول عناوین استنادی استفاده گردید. پیش از تشکیل این ماتریس به حذف واژگان موجود در سیاهه بازدارنده (با استفاده از سیاهه بازدارنده تعریف‌شده در این نرم‌افزار که با عنوان stopword (English) مشخص شده است) و ریشه‌یابی واژگان باقیمانده با استفاده از ریشه‌یاب پورتر[3] (Porter 1980) موجود در نرم‌افزار رپیدماینر پرداخته شد.

٤-٤. خوشه‌بندی فازی پروانه‌های ثبت اختراع

پس از به‌دست آوردن خصیصه‌های وزن داده‌شده در نرم‌افزار رپیدماینر، از برنامه

1. http://www.google.com/patents
2. http://www.freepatentsonline.com/search.html
3. Porter stemmer

نوشته‌شده به زبان پرل برای ایجاد خوشه‌بندی فازی c میانگین استفاده شده است. ازآنجاکه خوشه‌بندی مرجع دارای ۷۵ خوشه بود، در تنظیمات خوشه‌بندی، تعداد خوشه برابر با ۷۵ خوشه (طی ۱۰۰ تکرار) تعیین شد.

خروجی این برنامه جدولی است که در آن میزان درجه عضویت هر پروانه ثبت اختراع را به هریک از ۷۵ خوشه نشان می‌دهد.

جدول ۱. بخشی از ماتریس مربوط به عضویت فازی هر مدرک در هر خوشه براساس اشتراک استنادی

	A	T1	T2	T3	T4	T5	T6
2	T"""2155658"""	0.0133	0.0133	0.0133	0.0133	0.0133	0.0133
3	T"""3203319"""	0.0133	0.0133	0.0133	0.0133	0.0133	0.0133
4	T"""3406228"""	0.0133	0.0133	0.0133	0.0133	0.0133	0.0133
5	T"""3410880"""	0.013	0.013	0.013	0.013	0.013	0.013
6	T"""3450673"""	0.0133	0.0133	0.0133	0.0133	0.0133	0.0133
7	T"""3485806"""	0.0133	0.0133	0.0133	0.0133	0.0133	0.0133
8	T"""3488327"""	0.0133	0.0133	0.0133	0.0133	0.0133	0.0133
9	T"""3488389"""	0.0133	0.0133	0.0133	0.0133	0.0133	0.0133
10	T"""3499032"""	0.0133	0.0133	0.0133	0.0133	0.0133	0.0133

جدول ۱ نشان‌دهنده عضویت‌های فازی پروانه‌های ثبت اختراع در خوشه‌بندی حاصل از اشتراک در مآخذ است. در ستون A، شماره پروانه‌های ثبت اختراع و در ستون‌های بعدی، نام هر خوشه قرار دارد.

جدول ۲. بخشی از ماتریس مربوط به عضویت فازی هر مدرک در هر خوشه براساس واژگان عناوین استنادی

	A	T1	T2	T3	T4	T5	T6
2	T2155658	0.0133	0.0133	0.0133	0.0133	0.0133	0.0133
3	T3203319	0.0133	0.0133	0.0133	0.0133	0.0135	0.0133
4	T3406228	0.0133	0.0133	0.0131	0.0131	0.0132	0.0135
5	T3410880	0.013	0.0132	0.0131	0.013	0.013	0.0133
6	T3450673	0.0133	0.0133	0.0135	0.0133	0.0133	0.0134
7	T3485806	0.0128	0.0154	0.0127	0.0127	0.0128	0.0132
8	T3488327	0.0133	0.0133	0.0133	0.0133	0.0133	0.0133
9	T3488389	0.0133	0.0133	0.0133	0.0133	0.0133	0.0133
10	T3499032	0	0	0	0	0	0

در جدول ۲، عضویت‌های فازی پروانه‌های ثبت اختراع به هر خوشه در خوشه‌بندی مربوط به واژگان عناوین استنادها دیده می‌شود. در ستون A،شماره پروانه‌های ثبت اختراع و در ستون‌های بعدی، نام هر خوشه قرار گرفته است.

۵. یافته‌ها

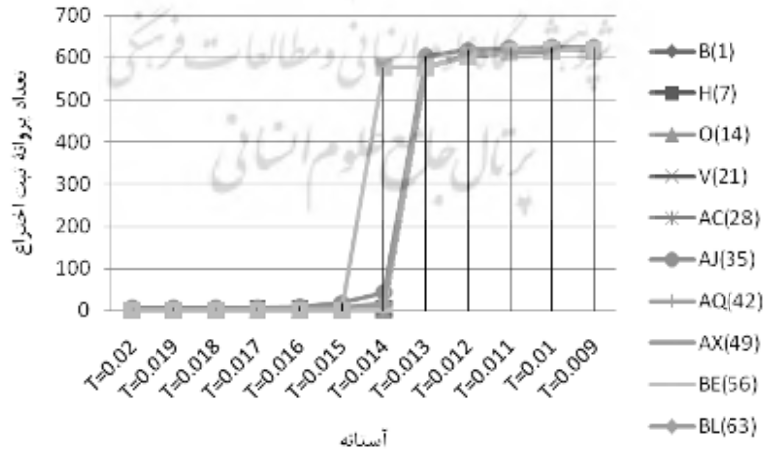
آمار مربوط به درجه عضویت پروانه‌های ثبت اختراع در هر خوشه، در خوشه‌بندی با استفاده از اشتراک در مآخذ و واژگان عناوین استنادها در جدول ۳ آمده است.

جدول ۳. درجه عضویت در هر خوشه - اشتراک در مآخذ و واژگان عناوین استنادها

	درجه عضویت در خوشه (واژگان عناوین استنادها)	درجه عضویت در خوشه (اشتراک در مآخذ)
Min	0	0
Max	1	1
Average	0.013332	0.013314
Median	0.0128	0.0133
Mode	0	0.0133
Std	0.041732	0.04051

پس از مشخص شدن درجه‌های عضویت هر پروانه ثبت اختراع، در ۷۵ خوشه آستانه‌ای برای ارزیابی مانعیت‌محور و آستانه‌ای برای ارزیابی جامعیت‌محور تعیین شده است. این آستانه با استفاده از مطالعه مقدماتی بر روی تعدادی از خوشه‌ها تعیین شده است. برای تعیین آستانه مناسب، آستانه‌های موجود در بازه ۰/۰۲ تا ۰/۹۰۰ مورد بررسی قرار گرفته‌اند. این بازه به‌طور تجربی و با توجه به شاخص‌های مربوط به میانگین و میانه و نیز انحراف معیار در جدول ۵ و ۶ انتخاب گردید.

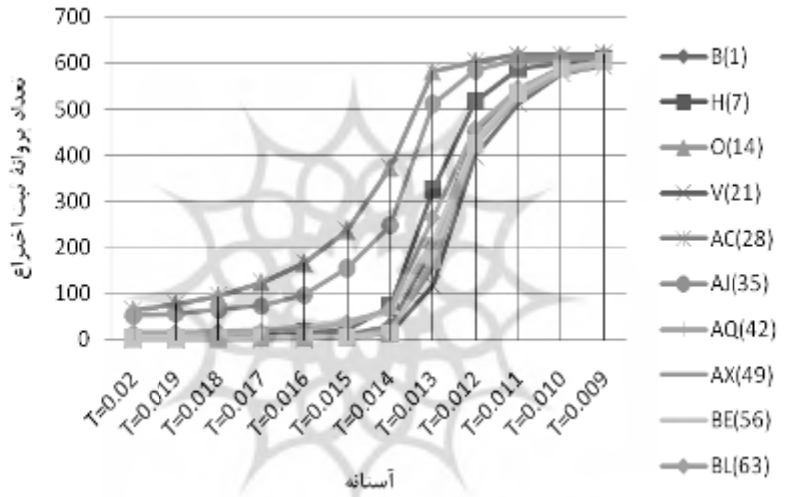
نمودارهای مربوط به تغییرات آستانه و تغییرات تعداد اعضای هر خوشه برای تعداد خوشه انتخابی در ادامه ارائه شده است.



نمودار ۱. اشتراک در مآخذ- تغییرات تعداد اعضای هر خوشه با توجه به تغییرات آستانه عضویت

مطابق نمودار ۱ کـه مربـوط بـه خوشـه‌بنـدی مبتنـی بـر اسـتناد (اشـتراک در مآخـذ) اسـت، تغییرات تعداد اعضای هر خوشه، از آستانه ۰/۰۱۵ به ۰/۰۱۳ شدت گرفته (شـیب منحنـی) و پـس از آستانه ۰/۰۱ تغییر چندانی نکرده است.

مطابق نمودار ۲ که مربوط بـه خوشـه‌بنـدی بـا اسـتفاده از عنـاوین اسـتنادی اسـت، تغییـرات تعداد اعضای هـر خوشـه در ایـن خوشـه‌بنـدی در آسـتانه‌ای میـان ۰/۰۱۲و ۰/۰۱۴ شـدت گرفتـه (شیب منحنی) و پس از آستانه ۰/۰۱ تغییر چندانی نکرده است.



نمودار ۲. واژگان عناوین استنادها- تغییرات اعضای هر خوشه با تغییر آستانه عضویت در هر خوشه

با توجه به نتایج حاصل از این مطالعه مقدماتی، در پژوهش حاضر در ارزیابی مانعیـت‌محـور از آستانه ۰/۰۱۴ و در ارزیابی جامعیت‌محور از آسـتانه ۰/۰۰۹ بـرای تعیـین اعضـای هـر خوشـه استفاده شده است و برای بررسی رونـد تغییـرات از مانعیـت‌محـوری بـه‌سـمت جامعیـت‌محـوری آستانه‌های ۰/۰۱۴، ۰/۰۱۳، ۰/۰۱۲، ۰/۰۱۱، ۰/۰۱۰ و ۰/۰۰۹ در تعیـین اعضـای هـر خوشـه مـورد استفاده قرار گرفته است.

در میان آسـتانه‌هـای تعیـین‌شـده، ارزیـابی صـورت‌گرفتـه پـس از اعمـال آسـتانه ۰/۰۱۴ در خوشه‌بندی با استفاده از واژگان عناوین استنادها هیچ نتیجه‌ای دربرنداشت. این بـدان معنـی اسـت که در این آستانه نقاطی که در خوشه‌بندی با یکدیگر هم‌رخداد شده‌انـد، در خوشـه‌بنـدی مرجـع هم‌رخداد نشده‌اند و برعکس. این امر سبب می‌شود که مخرج کسر در فرمول مورد اسـتفاده، هـم

در محاسبه مانعیت گسترش‌یافته (به‌دلیل عدم هم‌رخدادی در خوشه‌بندی مرجع) و هم در محاسبه جامعیت گسترش‌یافته (به‌دلیل عدم هم‌رخدادی در خوشه‌بندی) صفر شود و کسری تعریف‌نشده (به‌لحاظ ریاضی) حاصل گردد. مقادیر جامعیت و مانعیت گسترش یافته بی‌کیوبد با تغییر آستانه‌ها در جدول ۴ و نمودار ۳ نمایان است.

جدول ٤. مقادیر نهایی جامعیت و مانعیت گسترش‌یافته در خوشه‌بندی با استفاده از اشتراک در مآخذ و خوشه‌بندی با استفاده از واژگان عناوین استنادها در آستانه‌های اعمال‌شده

واژگان عناوین استنادها		اشتراک در مآخذ		
ج. گ. ب.	م. گ. ب.	ج. گ. ب.	م. گ. ب .	آستانه
_	_	0	1	0.014
0.26	1	0.7	1	0.013
0.52	1	0.73	1	0.012
0.63	1	0.75	1	0.011
0.70	1	0.75	1	0.010
0.73	1	0.76	1	0.009
0.568	1	0.615	1	میانگین
0.63	1	0.74	1	میانه

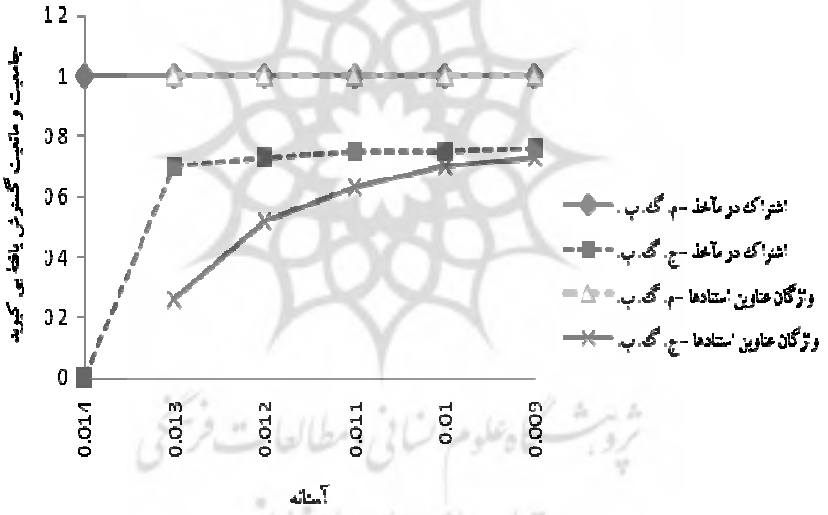
همان‌طور که جدول ۷ نشان می‌دهد، در آستانه ۰/۰۱۴ در خوشه‌بندی که براساس واژگان عناوین استنادی انجام شده است، جامعیت و مانعیت گسترش‌یافته نتیجه‌ای دربرنداشته است، بدین معنی که در این آستانه یکی از خوشه‌ها ازهم‌پاشیده شده و شاید چند گره (نود) باقی‌مانده باشد. در همین آستانه، در خوشه‌بندی با استفاده از اشتراک استنادی مقدار جامعیت گسترش‌یافته صفر است، اما مانعیت آن عدد ۱ را نمایش می‌دهد. همچنین، به‌ترتیب هرچه از آستانه ۰/۰۱۴ به‌سمت پایین‌تر سیر می‌شود، یعنی در آستانه‌های ۰/۰۱۳، ۰/۰۱۲، ۰/۰۱۱، ۰/۰۱۰، و ۰/۰۰۹ مقدار جامعیت گسترش‌یافته در هر دو خوشه‌بندی افزایش می‌یابد.

در آستانه‌های تعیین‌شده، میانگین میزان مانعیت گسترش‌یافته بی‌کیوبد در خوشه‌بندی پروانه‌های ثبت اختراع با استفاده از اشتراک در مآخذ برابر با ۱ و میانگین میزان جامعیت آن برابر با ۰/۶۱۵ است (پاسخ به پرسش ۱).

میانگین میزان مانعیت گسترش‌یافته بی‌کیوبد در خوشه‌بندی پروانه‌های ثبت اختراع با

استفاده از واژگان عناوین استنادها برابر با ۱ و میانگین میزان جامعیت آن برابر با ۰/۵۶۸ است (پاسخ به پرسش ۲).

مطابق با جدول ۷، همان‌طور که مشهود است گرچه هرچه از آستانه تعریف‌شده کاسته می‌شود مقدار جامعیت گسترش‌یافته بی‌کیوبد در هر دو خوشه‌بندی افزایش می‌یابد، نرخ تغییرات در اشتراک استنادی کمتر از خوشه‌بندی واژگان عناوین استنادهاست. از سوی دیگر، مقایسه ستون‌های مربوط به جامعیت در هر دو خوشه‌بندی نشان می‌دهد که میزان جامعیت در اشتراک استنادی بیش از خوشه‌بندی با استفاده از واژگان عناوین استنادهاست، اما میزان مانعیت در هر دو برابر با یک است (پاسخ به پرسش ۳). نمودار ۳ نشان‌دهنده تغییرات جامعیت و مانعیت گسترش‌یافته برحسب تغییر در آستانه و حرکت از جامعیت‌محوری در خوشه‌بندی به‌سمت مانعیت‌محوری است.



نمودار ۳. تغییرات جامعیت و مانعیت گسترش‌یافته برحسب تغییرات آستانه

با توجه به نمودار ۳، هرچه از مانعیت‌محوری (آستانه ۰/۰۱۴) در خوشه‌بندی به‌سمت جامعیت‌محوری (آستانه ۰/۰۰۹) پیش می‌رویم، میزان جامعیت گسترش‌یافته بی‌کیوبد در هر دو خوشه‌بندی افزایش می‌یابد. با این حال، در تمامی آستانه‌های تعیین‌شده میزان مانعیت گسترش‌یافته مقداری ثابت و برابر با یک را به‌خود اختصاص داده است.

مقایسه منحنی جامعیت گسترش‌یافته (ج.گ.ب.) در خوشه‌بندی حاصل از اشتراک در مآخذ با منحنی خوشه‌بندی حاصل از واژگان عناوین استنادها نشان می‌دهد که در یک آستانه

واحد، میزان جامعیت گسترش‌یافته در خوشه‌بندی با استفاده از اشتراک در مآخذ مقدار بیشتری را نسبت به خوشه‌بندی با استفاده از واژگان عناوین استنادها به‌خود اختصاص داده است.

همچنین، یافته‌های نمودار ۳ حاکی از آن است که خوشه‌بندی با استفاده از اشتراک در مآخذ در طیف گسترده‌تری از آستانه قابل تعریف است (پاسخ به پرسش ۳). این بدان معنی است که ساختار خوشه‌ای در خوشه‌بندی با استفاده از اشتراک در مآخذ از انسجام بیشتری نسبت به استفاده از واژگان عناوین استنادها برخوردار است.

٦. نتیجه‌گیری و بحث

یافته‌های پژوهش حاضر حاکی از آن است که خوشه‌بندی با استفاده از اشتراک در مآخذ در طیف گسترده‌تری از آستانه‌های تعیین‌شده (میزان جامعیت‌محوری) قابل تعریف است. این بدان معنی است که ساختار خوشه‌ای در خوشه‌بندی با استفاده از اشتراک در مآخذ از انسجام بیشتری نسبت به استفاده از واژگان عناوین استنادها برخوردار است. پژوهش سالتون حاکی از عدم برتری خصیصه استناد نسبت به خصیصه واژه در بازیابی مدارک بود (Salton 1963)، اما پژوهش حاضر حاکی از برتری خصیصه استناد است. این امر ممکن است ناشی از شیوه ایجاد استناد در پروانه‌های ثبت اختراع باشد. زیرا استنادها از سوی بازرسان و در ارتباط با ادعاهای اختراع ایجاد می‌شوند و همین امر، احتمال حضور منابع مرتبط در میان استنادها را افزایش می‌دهد. شاو نیز در پژوهش خود به این نتیجه رسید که بازنمایی استنادی، حضور ساختار خوشه‌ای را در طیف گسترده‌تری از سطوح مختلف جامعیت نسبت به بازنمایی موضوعی نشان داده است و بازنمایی‌های استنادی مورد استفاده در پژوهش او دارای ساختار خوشه‌ای در کمترین حالت جامعیت بوده است (Shaw 1990). سالتون نیز به این نتیجه رسید که اضافه کردن مفاهیم استنادی به توصیفگرهای موضوعی موجب افزایش مانعیت تا بیش از ۱۰ درصد در آستانه‌ای از جامعیت خواهد شد. همچنین، خصیصه استنادی عملکرد بسیار بهتری در سطوح پایین جامعیت‌محوری دارد (Salton 1971). شاو نیز به این نتیجه می‌رسد که در میان بازنمایی‌های ترکیبی جامعیت‌محور، مواردی که منابع استنادکننده و استنادشونده را مورد استفاده قرار داده‌اند نسبت به بازنمایی‌های حاصل از خصیصه‌های موضوعی، به‌طور قابل ملاحظه‌ای از اثربخشی بیشتری برخوردار بوده‌اند (Shaw 1991).

همچنین، نتایج پژوهش حاضر حاکی از برتری خصیصه استناد به‌صورت یک واحد (به‌دلیل برتری خوشه‌بندی حاصل از اشتراک در مآخذ) نسبت به تجزیه آن به‌صورت خصیصه واژه است. این نتیجه تا حدودی همگام با نتایج پژوهش فوجی است که در آن نمره‌دهی

استنادی عملکرد بهتری نسبت به نمرده دهی برمبنای واژه داشته است (Fujii 2007). با این حال، تیوانا و هورویتز بیان داشته‌اند که استفاده از استناد در الگوریتم پیشنهادی آنها سبب بازیابی تعداد زیادی از منابع نامرتبط در کنار منابع مرتبط گشته است (Tiwana and Horowitz 2009).

تجزیه استنادها به واژگان عناوین استنادها و اعمال پردازش‌هایی مانند حذف سیاهه بازدارنده و ریشه‌یابی واژگان، منجر به کاهش تعداد خصیصه‌های مورد استفاده در عمل خوشه‌بندی خواهد شد. در نتایج به‌دست‌آمده تعداد خصیصه استناد (در اشتراک در مآخذ) حدود ۵۰۰۰ و تعداد خصیصه واژگان عناوین استنادها حدود ۳۰۰۰ واژه بوده است. این امر نشان از برتری واژگان عناوین استنادها نسبت به استفاده صرف از استنادها، در کاهش حجم مورد استفاده برای ذخیره خصیصه‌ها دارد. با توجه به اینکه میزان جامعیت در خوشه‌بندی با استفاده از واژگان عناوین استنادها در آستانه‌های جامعیت‌محور نزدیک به میزان جامعیت در اشتراک در مآخذ است، هرگاه جامعیت از ارزش کمتری برخوردار باشد، به‌نظر می‌رسد استفاده از این خصیصه گزینه‌ای مناسب در خوشه‌بندی پروانه‌های ثبت اختراع باشد.

اما از سوی دیگر، جنبه حقوقی در ثبت اختراعات سبب شده است که نیازهای اطلاعاتی در جستجوی پروانه‌های ثبت اختراع نیازمند جامعیت بیشتر در نتایج جستجو باشد تا افرادی که در جهت بررسی قابلیت ثبت اثری به‌عنوان اختراع هستند بتوانند تمامی منابع مرتبط را مورد بررسی قرار دهند. از این رو، با توجه به نتایج پژوهش حاضر پیشنهاد می‌شود که روش اشتراک در مآخذ در طراحی نظام‌های بازیابی پروانه‌های ثبت اختراع مورد استفاده قرار گیرد.

۷. منابع

داروغه، شیرین. ۱۳۸۰. بررسی میزان همپوشانی کلیدواژه‌های عنوان، عناوین مأخذ با توصیفگرهای نمایه‌سازی در پایان‌نامه‌های دکترای تخصصی روان‌پزشکی، زنان و زایمان و قلب و عروق دانشگاه علوم پزشکی ایران. پایان‌نامه کارشناسی ارشد، دانشکده مدیریت و اطلاع‌رسانی پزشکی، دانشگاه علوم پزشکی و خدمات بهداشتی درمانی ایران.

Amigo, E., J. Gonzalo, J. Artiles, and F. Verdejo. 2008. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval* 12 (4): 461-486. http://www.springerlink.com/content/812x06387l52p045/fulltext.pdf (accessed 3 Jun. 2011).

Bagga, A. , and B. Baldwin. 1998. Algorithms for Scoring Coreference Chains. In *Proceedings of the Linguistic Coreference Workshop at the First International Conference on Language Resources and Evaluation (LREC'98),* 563-566. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.47.5848 (accessed 2 Feb. 2012).

Bonino,D., A. Ciaramella, and F. Corno. 2010. Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. *World Patent Information* 32 (1): 30 –38.

Class definition for class 977 - NANOTECHNOLOGY . 2011. http://www.uspto.gov/web/patents/classification/uspc977/defs977 .htm#C977S774000 (accessed 04 Sep. 2011).

Dhillon I., J. Kogan, and C. H. Nicholas. 2003. Feature selection and document clustering. In *A Comprehensive Survey of Text Mining*, 73-100. to be published by Springer-Verlag, M. Berry (Ed.). http://www.cs.umbc.edu/ csee/research/cadip/2002Symposium/kogan.pdf (accessed 29 Jul. 2011).

Fujii, A. 2007. Integrating content and citation information for the NTCIR-6 patent retrieval task. In *Proceedings of NTCIR-6 Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, 377-380. Tokyo, Japan,* 377-380. http://research.nii.ac.jp/ntcir/ntcir-ws6/ OnlineProceedings/NTCIR/76.pdf (Accessed 22 April 2012).

Generation of a test collection based on citations. [n.d.]. http://www.ir-facility.org/citation-based-test-collection (accessed 04 Sep. 2011).

Graf, E., and L. Azzopardi. 2008. A methodology for building a patent test collection for prior art search. In *Proceedings of the 2nd International Workshop onEvaluating Information Access (EVIA)*, December 16, 2008, Tokyo, Japan, 60–71. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/EVIA2008/11-EVIA2008-GrafE.pdf (accessed 27 Jan. 2010).

Hideo, J., L. Azzopardi, and W. Vanderbauwhede. 2010. A survey of patent users. In *Proceeding of the third symposium on Information interaction in context, August 18-22,* 13-22. New Brunswick, New Jersey, USA: ACM. doi> 10.1145/1840784.1840789.

Huang, A. 2008. Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference NZCSRSC 2008 Christchurch New Zealand*, 49-56. http://nzcsrsc08.canterbury.ac.nz/site/ proceedings/Individual_Papers/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf (accessed 30 May 2011).

Jain, A. K., and R. C. Dubes. 1988. Algorithms for clustering data. Prentice-Hall advanced reference series. NJ: Prentice-Hall, Inc., Upper Saddle River.

Jain, A., M. Murty, and P. Flynn. 1999. Data clustering: A review, *ACM Computing Surveys* 31 (3):264 – 323. doi>10.1.1.18.2720.

Kang, I., S. Na, J. Kim, and J. Lee. 2007. Cluster-based patent retrieval. *Information Processing and Management* 43 (5): 1173-1182.

Kessler, M. M. 1963. An experimental study of bibliographic coupling between technical papers. *IEEE transactions on information theory* 9 (1): 49-51.

Lai K. K., and S. H. J. Wu. 2005. Using the patent co-citation approach to establish a new patent classification system. *Information Processing and Management* 41 (2): 313–330.

Larsen, B., and C. Aone. 1999. Fast and effective text mining using linear-time document clustering. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 16-22. San Diego, California: ACM . doi>10.1145/312129.312186.

Leydesdorff, L. 1987. Various Methods for the mapping of science. *Scientometrics* 11 (5-6): 295-324.

Li, X., H. Chen, Z. Zhang, and J. Li. 2007. Automatic patent classification using citation network information: An experimental study in nanotechnology. In *JCDL'07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, 419–427. New York, NY, USA. ACM. doi>10.1145/1255175.1255262.

Porter, M. F. 1980. An algorithm for suffix stripping. *Program: Electronic Library and Information Systems* 14 (3): 130-137. http://telemat.det.unifi.it/book/2001/wchange/download/stem_porter.html (accessed 20 Jun. 2011).

Salton, G. 1963. Associative document retrieval techniques using bibliographic information. *Journal of the ACM* 10 (4): 440-457. http://delivery.acm.org/10.1145/330000/321188/p440-salton.pdf?ip=217.218.83.9&CFID=36477701&CFTOKEN =92443976&__acm__=1313915650_9ed0bb61c9157adf2e126a72446f5244 (accessed 22 Aug. 2011). doi> 10.1145/321186.321188.

Salton, G. 1971. Automatic indexing using bibliographic citations. *Journal of documentation* 27 (2): 98-110.

Shaw, W. M., Jr. 1990. Subject indexing and citation indexing. part I: Clustering structure in the cystic fibrosis document collection. *Information Processing and Management* 26 (6): 693-703.

Shaw, W. M., Jr. 1991. Subject and citation indexing. Part II: The Optimal, cluster-based retrieval performance of composite representations. *Journal of The American Society For Information Science* 42 (9):676-664.

tf–idf. 2011. Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Tf%E2%80%93idf (accessed 1 Jun. 2011).

Tan, P. N., M., Steinbach, and V. Kumar. 2006. Cluster analysis: Basic concept and algorithm. In *Introduction to Data Mining*. P. N. Tan, M. Steinbach and V. Kumar, 487-568. Boston, MA: Addison-Wesley Longman Publishing.

Tiwana, S., and E. Horowitz. 2009. Find cite: automatically finding prior art patents. In *Proceedings of the 2nd international workshop on Patent information retrieval*, 37-40. http://delivery.acm.org/10.1145/1660000/1651352/p37-tiwana.pdf?ip=217.218.83.9&CFID=39566988&CFTOKEN=46635715&__acm__=1315641819_3628c1ae1982ae3f1f1f9b5747a3f756 (accessed 3 Mar. 2011).

United States Patent and Trademark Office (USPTO). 2010. Manual of Patent Examining Procedure (MPEP) (eighth edition). http://www.uspto.gov/web/offices/pac/mpep/pdf_download_search_instructions.doc (accessed 9 Jan. 2011).

Wedding, D. K. 2009. Extending the Data Mining Software Packages SAS enterprise miner and spss clementine to handle fuzzy cluster membership: Implementation with examples (Master of Science Thesis in Data Mining), Central Connecticut State University. http://web.ccsu.edu/datamining/Data%20Mining%20Theses/Don%20Wedding%20thesis.pdf (accessed 13 May 2011).

Xue, X., and W. B. Croft. 2009. Transforming patents into prior-art queries. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 808-809. Boston, MA, USA. ACM. doi>10.1145/1571941.1572139.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

Application of Bibliographic Coupling versus Cited Titles Words in Patent Fuzzy Clustering

Anahita Kermani*
Master in Library and Information Science

Narges Neshat[1]
Associated Professor in National Library and Archives of Iran

Abbas Horri[2]
Professor in Library and Information Science, Tehran University

Iranian Journal of Information Processing & Management
Iranian Research Institute For Science and Technology
ISSN 2251-8223
eISSN 2251-8231
Indexed in LISA, SCOPUS & ISC
Vol.28 | No.2 | pp: 411-432
Winter 2013

Abstract: Attribute selection is one of the steps before patent clustering. Various attributes can be used for clustering. In this study, the effect of using citation and citation title words, respectively, in form of bibliographic coupling and citation title words sharing, were measured and compared with each other, as patent attributes. This study was done in an experimental method, on a collection of 717 US Patent cited in the patents belong to 977/774 subclass of US Patent Classification. Fuzzy C-means was used for patent clustering and extended BCubed precision and extended BCubed recall were used as evaluation measure. The results showed that the clustering produced by bibliographic coupling had better performance than clustering used citation title words and existence of cluster structure were in a wider range of exhaustivity than citation title words.

Keywords: patents, patent clustering, bibliographic coupling, citation title words, Fuzzy C-means clustering, clustering evaluation, BCubed Precision, BCubedRecall

*Corresponding author: kermanianahita@gmail.com
1. narges_neshat@yahoo.com 2. riwash@yahoo.com