بررسی عناصر و مؤلفه‌های رابط کاربر در نظام‌های بازیابی اطلاعات مبتنی بر خوشه‌بندی[1]

سید مهدی حسینی *

دانشجوی دکتری علوم کتابداری و اطلاع‌رسانی،
دانشگاه فردوسی مشهد

دریافت: ۱۳۸۸/۱۰/۱۲ | پذیرش: ۱۳۸۹/۰۲/۲۸

فصلنامهٔ علمی پژوهشی
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا(چاپی) ۵۲۰۶-۱۷۳۵
شاپا(الکترونیکی) ۵۵۸۳-۲۰۰۸
نمایه در LISA، SCOPUS و ISC
http://jist.irandoc.ac.ir
دوره ۲۶ | شمارهٔ ۳ | ص ص ۶۲۵ - ۶۵۳
بهار ۱۳۹۰

نوع مقاله: پژوهشی

چکیده: موتورهای خوشه‌بندی، به‌موازات دو روش جستجوی کلیدواژه‌ای و مروری، برای کمک به بازیابی اطلاعات مرتبط موجود در وب، در سطح وسیعی، به‌کار می‌روند. در این نظام‌ها، رابط کاربر، به‌عنوان نقطه تماس با کاربران و برای استفاده از بازخورد در بهبود پرسش کاربر، از اهمیت زیادی برخوردار است. هدف این مقاله، شناسایی عناصر و مؤلفه‌های مهم رابط کاربر نظام‌های خوشه‌بندی است. ابتدا، در مورد ضرورت استفاده از خوشه‌بندی در نظام‌های بازیابی اطلاعات، مبانی نظری، الگوریتم‌ها، کاربردها، و معیارهای ارزیابی خوشه‌بندی بحث شده است. سپس، ضمن شناسایی ۱۰ موتور خوشه‌بندی بیان شده در متون حوزه بازیابی اطلاعات در وب، عناصر و مؤلفه‌های رابط کاربر این‌گونه نظام‌ها و عملکرد هر کدام، به‌تفکیک، توضیح داده شده است. در پایان، میزان استفاده از عناصر و مؤلفه‌های شناسایی شده در موتورهای خوشه‌بندی مورد مطالعه، بیان شده است.

کلیدواژه‌ها: خوشه‌بندی، نظام‌های بازیابی اطلاعات، رابط کاربر، نظام‌های خوشه‌بندی، موتورهای جستجو

*. smahdihosseni@gmail.com

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

۱. نویسنده سپاسگزاری قلبی خود را از استاد محترم دکتر رحمت‌الله فتاحی به خاطر رهنمودهای ارزنده‌شان در تدوین این مقاله ابراز می‌دارد.

۱. مقدمه

جان نایسبت[1] معتقد است: «ما در اطلاعات غوطه‌وریم، اما از کمبود دانش رنج می‌بریم.» (در Dietinger et al. 1999) شاید به‌همین دلیل است که دایتینگر و همکارانش وجود مخازن دانش بسیار بزرگ، بدون ساختار و افزایش سریع اطلاعات را دو ویژگی جوامع اطلاعاتی کنونی می‌دانند.

نظام‌های بازیابی اطلاعات دسترسی به مجموعه عظیمی از منابع اطلاعاتی را برای تعداد زیادی از کاربران فراهم می‌کنند. به‌طور معمول، مجموعه‌های اولیه منابع وب از اسناد متنی تشکیل شده بود، اما امروزه، حجم انواع جدیدی از داده‌ها مانند تصویر، صوت، ویدیو، و غیره در حال افزایش است. در ابتدا، فعالیت‌هایی مانند تحلیل محتوا و نمایه‌سازی و رده‌بندی اسناد توسط متخصصان موضوعی یا نمایه‌سازان آموزش دیده، به‌صورت دستی انجام می‌شد، ولی اتکا به نیروی انسانی در توسعه و نگهداری چنین مجموعه‌هایی پرهزینه و زمان‌بر و مستعد خطاست (Cunningham et al. 1998). به‌تدریج، قابلیت خواندن محتوای اسناد با ماشین، همراه با افزایش قدرت پردازش رایانه‌های جدید، امکان نمایه‌سازی خودکار را به‌طور کامل فراهم نموده است. از این رو، بسیاری از فرایندهای دستی بازیابی اطلاعات را می‌توان با روش‌های هوشمند جایگزین نمود. با افزایش حجم اسنادی که قابلیت خواندن با ماشین را دارند و رشد روند ایجاد کتابخانه دیجیتالی، اهمیت کاربرد چنین نظام‌هایی دو چندان شده است (Cunningham et al. 1998).

به‌طور کلی، هدف اصلی نظام‌های بازیابی اطلاعات، فراهم کردن امکان دسترسی کاربران به منابع مرتبط و متناسب در کوتاه‌ترین زمان ممکن است. به‌طور معمول، این نظام‌ها سعی دارند به دو شیوه، منابع مرتبط را با نیازهای کاربران منطبق نمایند: ۱) از طریق امکان جستجو با تعیین عبارت‌های جستجو[2] است که کاربر/ جستجوگر آنها را با نیاز اطلاعاتی مرتبط می‌داند؛ ۲) از طریق مرور[3] موضوعاتی که توسط متخصصان موضوعی دسته‌بندی و سازماندهی می‌شود. به‌طور معمول، کاربر زمانی از روش جستجو استفاده می‌کند که نیاز اطلاعاتی مشخصی دارد و می‌داند دنبال چه می‌گردد. به‌همین دلیل، آن را در قالب عبارت‌هایی که فکر می‌کند مفهوم نیاز اطلاعاتی وی را به اسناد ربط می‌دهد، به‌عنوان پرسش به نظام ارائه می‌دهد و نظام از طریق دستورالعمل‌هایی که برای تطابق به‌کار می‌گیرد، اسناد مرتبط را بازیابی نموده و به‌صورت رتبه‌بندی، فهرست آنها را در اختیار کاربر قرار می‌دهد. این رویکرد، تحول بزرگی در بازیابی اطلاعات ایجاد نمود، چون کاربر در مدت کوتاهی می‌توانست متوجه شود اسناد مرتبط با پرسش وی در پایگاه آثار وجود دارد یا خیر. البته، این روش مستلزم وجود واژگان و اصطلاحات مشترک بین کاربر و نظام برای

1. John Naisbett 2. search queries 3. browsing

بیان مفهوم مورد نظر بود. به‌دلیل تفاوت سطح دانش، زمینه ذهنی قبلی از موضوع، تجارب، و زبان کاربر با نویسندگان مقالات یا افرادی که بازنمون‌های اسناد را تهیه می‌نمایند، تطابق عبارت‌های جستجوی کاربر با کلیدواژه‌ها یا عبارت‌های به‌کار رفته در متن آثار، همواره یکی از مشکلات نظام‌های بازیابی اطلاعات و یافتن اسناد مرتبط با پرسش، با استفاده از روش جستجوی کلیدواژه‌ای، بوده است.

از سوی دیگر، گاهی کاربر به‌دنبال نیاز اطلاعاتی مشخصی نیست و فقط به‌منظور کسب آگاهی از آخرین تحولات حوزه موضوعی یا مرور آخرین نوشته‌های زمینه مورد علاقه‌اش به‌سراغ نظام بازیابی می‌رود. برای برآورده نمودن چنین نیازهایی، برخی نظام‌ها با استفاده از افراد آموزش‌دیده در زمینه سازماندهای اطلاعات یا متخصصان موضوعی، مطالب موجود در نظام را به‌روش‌های مختلفی مانند سلسله مراتبی یا الفبایی نمایش می‌دهند تا کاربر از طریق مرور بتواند به‌سمت مطالب مورد نظر هدایت شود. این رویکرد، اگر چه به‌دلیل استفاده از متخصصان موضوعی از دقت بیشتری برخوردار است، کند و پرهزینه و دامنه محتوای آن اندک است. همچنین، امکان دارد افراد مبتدی هنگام مرور، پس از چند مرحله دنبال کردن پیوندها، سردرگم شوند (Baeza-Yates and Riberio-Neto 1999).

ویژگی‌های محیط اطلاعاتی، مانند حجم بسیار زیاد منابع و وجود قالب‌های مختلف و پویایی، سبب شد که دو روش اشاره شده، در بسیاری از زمینه‌ها، قابلیت همگامی و تحت پوشش قرار دادن بخش وسیعی از اطلاعات را نداشته باشند از جمله: تصاویر و محتوای آنها، اخبار، قالب اطلاعات موجود در وب، تفکیک اصطلاحات دارای کاربرد در حوزه‌های مختلف، اشتراک لفظی/معنایی، مترادف‌ها، و رتبه‌بندی نتایج جستجو.

روش سومی که به‌منظور حل برخی از این مشکلات، مورد توجه قرار گرفت، خوشه‌بندی[1] بود. این روش، ترکیبی از بهترین ویژگی‌های جستجوی مبتنی بر پرسش و مبتنی بر گروه‌بندی اسناد است که در آن کاربر ممکن است با استفاده از تعریف ناقص پرسش روی یک موضوع کلی تمرکز نماید، سپس از طریق نتایج پرسش که به‌صورت موضوع‌های بسیار خاص و پویا ایجاد شده است، به‌طور عمیق، به جستجوی گروه پاسخ مرتبط بپردازد.

در سال‌های اخیر، نظام (موتور)های خوشه‌بندی نتایج جستجوی وب شهرت به نسبت عام، یافته‌اند. به احتمال، اولین موتور کاوش خوشه‌بندی تجاری، Northern Light بود که در اواخر دهه ۱۹۹۰ ارائه شد. یکی از پیشرفت‌های مهم در این زمینه، ارائه Vivisimo است که نتایج جستجو را در قالب خوشه‌های پویا نمایش می‌دهد. این نظام از سال ۲۰۰۱ تا ۲۰۰۳ جایزه بهترین ابرموتور کاوش را

1. clustering

که توسط SearchEngineWatch.com تهیه شده بود، از آن خود نمود. ویژگی مشترک موتورهای خوشه‌بندی کنونی این است که مانند ابرموتورهای کاوش از نمایه و نتایج جستجوی موتورهای کاوشی که به‌طور عمومی، دسترس پذیرند، استفاده می‌نمایند (Carpineto et al. 2009).

خوشه‌بندی نتایج جستجو، نه تنها مورد توجه شرکت‌های تجاری قرار گرفته است، یک حوزه فعال پژوهشی است که در آن، موضوعات و نظام‌ها را به‌صورت خاص، مورد بحث قرار داده‌اند. خوشه‌بندی نتایج جستجو، به‌طور مشخص با حوزه خوشه‌بندی اسناد ارتباط دارد.

موتور خوشه‌بندی تلاش می‌کند محدودیت‌های موتورهای کاوش کنونی را از طریق خوشه‌بندی نتایج، به‌عنوان یک ویژگی افزوده به رابط کاربر استاندارد خود، رفع نماید. تأکید می‌شود که موتور خوشه‌بندی بدلی برای موتورهای کاوش نیست، بلکه به‌عنوان تکمیل‌کننده آنها مد نظر است. برخی جنبه‌های جستجو و بازیابی که نشان می‌دهد موتور خوشه‌بندی می‌تواند بیشترین فایده را در تکمیل خروجی موتور جستجوی ساده ارائه دهد عبارتند از:

- *بازیابی سریع موضوع‌های فرعی*. اگر اسنادی که مربوط به یک موضوع فرعی هستند، به‌درستی درون یک خوشه قرار گرفته باشند و کاربر قادر باشد مسیر درست را از طریق برچسب خوشه انتخاب کند، می‌تواند اسناد را به‌جای جستجوی خط به خط، به‌صورت لگاریتمی (در مدتی کوتاه) جستجو نماید.
- *کاوش موضوعی*. یک سلسله مراتب از خوشه‌ها دیدگاهی وسیع از کل موضوع پرسش شامل اصطلاحاتی برای فرمول‌بندی مجدد پرسش فراهم می‌نماید که به‌ویژه، برای جستجوهای فرمول‌بندی مجدد در حوزه‌های ناشناخته یا پویا مفید است.
- *کاهش نادیده گرفتن اطلاعات*. کاربران وب، به‌طور معمول، فقط به اولین صفحه نتایج جستجو می‌نگرند، بنابراین بیشتر اطلاعات بازیابی شده که در صفحه اول امکان نمایش نمی‌یابد، نادیده گرفته می‌شود. از آنجا که موتور خوشه‌بندی، محتوای بسیاری از نتایج جستجو را در یک نگاه کلی در صفحه اول نتایج خلاصه می‌نماید، ممکن است کاربر صدها نتیجه مرتبط را بدون نیاز به بارگذاری و یا پیمایش صفحات فرعی مرور نماید (Carpineto et al. 2009).
- *کاهش شگرف زمان پردازش*. بدین معنی که حجم وسیعی از نمایه‌های چندین موتور جستجوی عمده و قوی را در زمان کوتاهی مورد کاوش قرار می‌دهد.
- *فراهم نمودن زمینه فهم بهتر موضوع*. برخلاف تصور اولیه کاربر، ممکن است پرسش وی دارای جنبه‌ها و معانی گوناگون باشد. نظام خوشه‌بندی، با ارائه جنبه‌ها یا معانی مختلف عبارات جستجو در ساختاری خوشه‌ای، به درک بهتر کاربر کمک می‌کند.

- *فراهم نمودن امکان کاوش نظام‌مند نتایج جستجو*. در این گونه نظام‌ها، نتایج جستجو در خوشه‌های مرتبط به هم ارائه می‌گردد. این ساختار به کاربر کمک می‌کند کنترل بیشتری بر جستجوی خود داشته باشد و نتایج جستجو را به‌ترتیب، از موضوعات کلی به موضوعات جزئی و برعکس، بررسی نماید.

۲. فرضیه خوشه‌بندی

بنیادی‌ترین هدف نظام بازیابی اطلاعات، یافتن منابعی است که با نیاز اطلاعاتی کاربر ربط بیشتری داشته باشد (Saracevic, 2007a, 2007b). ربط، مفهومی نسبی است و می‌تواند از ابعاد مختلفی مورد نظر باشد (Borlund and Ingwersen 1998; Mizzaro 1997; Saracevic 1999؛ فتاحی ۱۳۸۳). همچنین، اسناد و آثار اطلاعاتی دارای خصوصیات و ویژگی‌های متعدد از جنبه‌های گوناگون دارند (ارسطوپور ۱۳۸۸). با توجه به هر بعد و خصوصیاتی که برای پاسخگویی به گروهی از نیازهای اطلاعاتی ممکن است مناسب باشد، می‌توان اسناد مشابه را در کنار هم و در یک خوشه قرار داد.

فرضیه خوشه‌بندی نشان‌دهنده پیش‌فرض‌های بنیادی است که خوشه‌بندی در بازیابی اطلاعات بر اساس آنها استفاده می‌شود. فرضیه خوشه‌بندی به این صورت است: «اسنادی که در یک خوشه قرار می‌گیرند با توجه به نیاز اطلاعاتی مرتبط، رفتار (پاسخ) مشابهی دارند». به عبارت دیگر، این فرضیه می‌گوید اگر سندی از یک خوشه با درخواست جستجو ربط داشته باشد، سایر اسناد همان خوشه نیز با آن مرتبط هستند، زیرا خوشه‌بندی، اسنادی را در یک خوشه قرار می‌دهد که دارای اصطلاحات مشترک فراوانی هستند. فرضیه خوشه، به‌طور اساسی، فرضیه همجواری[1] است. در هر دو مورد، ما فرض می‌کنیم اسناد مشابه با توجه به ربط، رفتار مشابهی دارند (Manning, Raghavan, and Schutze 2008).

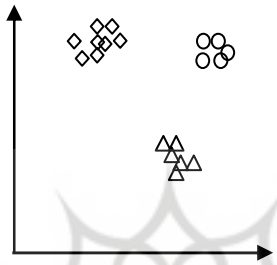
خوشه‌بندی، معمول‌ترین شکل یادگیری بدون نظارت[2] است (Cunningham et al. 1998)، یعنی خبرگی انسان در تعیین اسناد برای طبقه‌ها دخالت ندارد. در خوشه‌بندی اسناد، آرایش و توزیع داده‌ها ملاک عضویت در خوشه است. این عمل، بر اساس فرایندهای خاصی که الگوریتم خوشه‌بندی نامیده می‌شود، صورت می‌گیرد.

۳. الگوریتم‌های خوشه‌بندی

الگوریتم‌های خوشه‌بندی مجموعه‌ای از اسناد را در گروه‌های فرعی به‌نام خوشه، دسته‌بندی می‌کنند. هدف این است که اسناد درون یک خوشه با هم دارای بیشترین شباهت باشند، اما هر خوشه با خوشه‌های دیگر از کمترین شباهت برخوردار باشد. ورودی اصلی یک الگوریتم

1. contiguity 2. unsupervised learning

خوشه‌بندی، اندازه فاصله[1] است. در شکل ۱، اندازه فاصله از طریق سطح دوبعدی نشان داده شده است که این اندازه‌گیری، سه خوشه متفاوت ایجاد نموده است. در خوشه‌بندی اسناد، اندازه فاصله اغلب بر اساس فاصله اقلیدسی است. اندازه‌های فاصله متفاوت سبب ایجاد خوشه‌بندی متفاوت می‌شود، بنابراین اندازه‌گیری فاصله، ابزار مهمی است که از طریق آن می‌توان خوشه‌بندی نهایی را تحت تأثیر قرار داد (Manning, Raghavan, and Schutze 2008).



شکل ۱. مجموعه‌ای داده دارای ساختارخوشه‌بندی روشن

الگوریتم‌های خوشه‌بندی را می‌توان در دو دسته کلی قرار داد: ۱) الگوریتم‌هایی که مبنای آماری دارند، مانند استفاده از روش بردار فضایی و مدل‌های شبکه عصبی مصنوعی و مدل‌های احتمالاتی؛ ۲) الگوریتم‌هایی که بر مبنای ویژگی‌های اسناد عمل می‌کنند، مانند ویژگی‌های زبانی و فراداده و استنادها[2].

دایتنگر و همکارانش معتقدند که گروه‌بندی ممکن است جامع[3] و مانع[4] باشد یا از یک بازنمون غنی‌تر مانند سلسله مراتبی یا گروه‌های همپوشان تشکیل شده باشد. آنها برای ایجاد خوشه پیشنهاد نمودند از فراداده استفاده شود که به‌منظور تقویت اثربخشی و ایجاد روابط چند سویه باید سه ویژگی را از طریق فراداده‌ها دنبال نمود: ۱) **نوع** که بیانگر نوع رابطه بین دو گروه است مانند مفاهیم اصلی و فرعی، علت و معلول، متضاد و مترادف، پیش نیازها، معرف، و غیره؛ ۲) **ارزش وزنی** که مشخص‌کننده درجه کاربرد یک نوع خاص برای رابطه‌ای مشخص است، مانند ارزش‌های فازی از جمله عالی، خوب، متوسط، بد، و غیره که با درصد مشخصی بیان می‌شود؛ ۳) **ارزش کیفی** که بیانگر پایایی رابطه است و با درصد نشان داده می‌شود (Dietinger et al. 1999).

---

1. distance measure

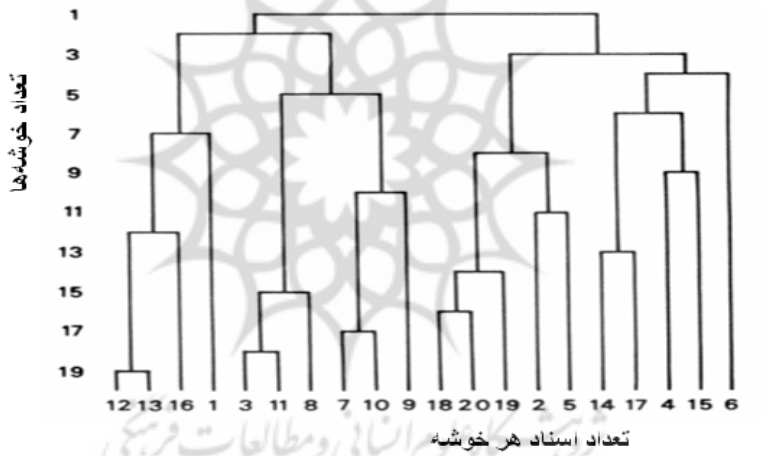
۲. جهت مطالعه مفصل انواع الگوریتم‌های خوشه‌بندی به مقاله‌های مروری Xu and Wunsch II 2005; Carpineto et al. 2009 و جهت مطالعه نمونه‌ای از کاربرد ویژگی‌های اسناد، به‌ویژه دوبلین کور، به مقاله ارسطوپور، ۱۳۸۸ نگاه کنید.

3. exhaustive 4. exclusive

٤. انواع خوشه‌بندی

خوشه‌بندی مسطح[1] (شکل ۱)، مجموعه‌ای مسطح از خوشه‌ها را بدون ساختار روشنی که خوشه‌ها را به همدیگر مرتبط سازد، ارائه می‌دهد. خوشه‌بندی سلسله مراتبی[2] (شکل ۲) خوشه‌ها را در ساختاری سلسله مراتبی ارائه می‌دهد.

دومین تمایز مهم بین الگوریتم‌های خوشه‌بندی سخت و نرم برقرار است. خوشه‌بندی سخت[3] یک تقسیم‌بندی بدون انعطاف از اسناد ارائه می‌دهد، یعنی هر سند می‌تواند فقط عضو یک خوشه باشد. تخصیص الگوریتم خوشه‌بندی نرم[4] انعطاف‌پذیراست، یعنی تخصیص هر سند، توزیعی روی تمامی خوشه‌هاست. در این تخصیص، هر سند در چندین خوشه عضویت دارد، اما رتبه عضویت آن ممکن است در هر خوشه متفاوت باشد. نمایه‌سازی معنایی پنهان[5] نمونه‌ای از الگوریتم خوشه‌بندی نرم است (Manning, Raghavan, and Schutze 2008).



شکل۲: ساختار خوشه‌بندی سلسله مراتبی

در خوشه‌بندی سخت دارای ساختار مسطح، هر سند فقط عضو یک خوشه است، اما در ساختار سلسله مراتبی، امکان عضویت کامل بیش از یک خوشه وجود دارد. به عبارت دیگر، در خوشه‌بندی چند سطحی[6] سلسله مراتبی تمامی اعضای یک خوشه عضو خوشه لایه بالاتر از خود هستند. تفاوت خوشه‌بندی سخت که امکان عضویت چندگانه را فراهم می‌کند با خوشه‌بندی نرم این است که در خوشه‌بندی سخت مقادیر یا ۰ است یا ۱، در حالی‌که در خوشه‌بندی نرم می‌توانند غیر از عدد منفی هر مقداری را داشته باشند.

1. flat clustering
2.hierarchical clustering
3. hard clustering
4. soft clustering
5. latent semantic indexing
6. partitional clustering

برخی پژوهشگران بین خوشه‌بندی جامع[1] که هر سندی را عضوی از یک خوشه می‌داند و خوشه‌بندی غیر جامع که در آن، برخی اسناد عضو هیچ خوشه‌ای نیستند، تمایز قایل می‌شوند. خوشه‌بندی غیر جامع که در آن یک سند عضو هیچ خوشه‌ای نیست یا عضو یک خوشه است، خوشه‌بندی انحصاری[2] نامیده می‌شود.

عمل خوشه‌بندی اسناد به‌طور کلی، دو نوع است: سراسری[3] و محلی[4]. در راهبرد خوشه‌بندی سراسری، اسناد به‌نسبت بسامدشان در کل مجموعه، دسته‌بندی می‌شوند. در این روش، تمامی اسناد مجموعه برای تعیین ساختار عمومی اصطلاحنامه مانندی که تعیین‌کننده روابط اصطلاحات است، به‌کار می‌رود. سپس، این ساختار به کاربر ارائه می‌شود تا اصطلاحات مورد نیاز برای گسترش جستجو را انتخاب نماید[5]. در راهبرد خوشه‌بندی محلی، گروه‌بندی اسناد متأثر از بافت تعریف شده توسط پرسش جاری و مجموعه اسناد بازیابی شده است (Baeza-Yates and Riberio-Neto 1999).

۵. کاربردهای خوشه‌بندی

جدول ۱ برخی از کاربردهای اساسی خوشه‌بندی را در بازیابی اطلاعات نشان می‌دهد که از نظر مجموعه اسنادی که خوشه‌بندی می‌کنند (نتایج جستجو، مجموعه یا زیرگروه‌های مجموعه) و جنبه‌هایی از نظام بازیابی اطلاعات که مورد توجه قرار می‌دهند (مانند تجارب کاربر، رابط کاربر، کارآمدی، و اثربخشی نظام جستجو) با هم فرق دارند. اما، همه آنها بر پیش‌فرض اصلی که توسط فرضیه خوشه‌بندی بیان شد، استوارند (Manning, Raghavan, and Schutze 2008).

جدول ۱. برخی از کاربردهای خوشه‌بندی در بازیابی اطلاعات
(Manning, Raghavan, and Schutze 2008)

کاربرد	چه چیز خوشه‌بندی می‌شود؟	فواید	نمونه
خوشه‌بندی نتایج جستجو	نتایج جستجو	ارائه اطلاعات مؤثرتر به کاربر	شکل ۳
پخش- جمع کردن (Scatter-gather)	زیرمجموعه‌های مجموعه	رابط کاربر بدیل: "جستجو بدون تایپ کلیدواژه"	شکل ۴

←

1. exhaustive 2.. exclusive 3. global 4. local

۵. برای نمونه به موتور خوشه‌بندی ask.com نگاه کنید.

ادامهٔ جدول ۱. برخی از کاربردهای خوشه‌بندی در بازیابی اطلاعات
(Manning, Raghavan, and Schutze 2008)

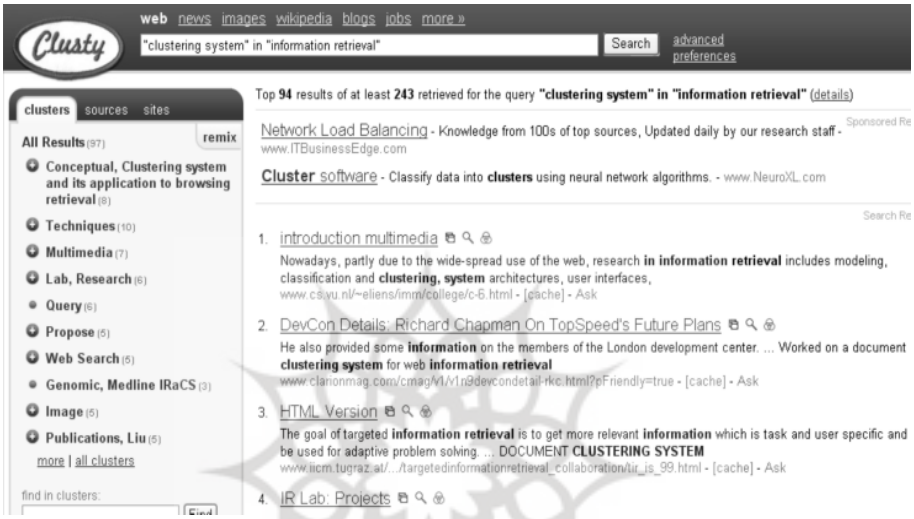
کاربرد	چه چیز خوشه‌بندی می‌شود؟	فواید	نمونه
خوشه‌بندی مجموعه	مجموعه	ارائه اطلاعات مؤثر برای مرورهای کاوشی	Mckeown et al. (2000) http://news.google.com
مدل‌سازی زبان	مجموعه	افزایش دقت و یا بازیافت	Liu and Craft (2004)
بازیابی مبتنی بر خوشه	مجموعه	کارآمدی بالا: جستجوی سریع	Salton (1971a)

۵-۱. کاربرد اول: خوشه‌بندی نتایج جستجو

روش‌های خوشه‌بندی، بیشتر در بازیابی اطلاعات برای تبدیل پرسش اولیه به‌شکلی بهتر و به‌منظور بازنمایی نیاز اطلاعاتی کاربر، صورت می‌گیرد. از این دیدگاه، خوشه‌بندی، بیشتر عملی است که پرسش کاربر را تبدیل می‌کند نه متن اسناد را (Baeza-Yates and Riberio-Neto 1999). خوشه‌بندی نتایج جستجو، به‌منظور تکمیل چرخه بازخورد ربط توسط کاربر، انجام می‌شود. کاربر اسناد بازیابی شده را بررسی و اسناد مرتبط را مشخص می‌نماید، سپس اطلاعات حاصل از رفتار کاربر برای انتخاب اصطلاحات جدید و بسط جستجو توسط نظام به‌کار می‌رود. استدلال این است که پرسش بسط یافته سبب بازیابی بیشتر اسناد مرتبط خواهد شد. بنابراین، ایده بنیادی استفاده از خوشه‌بندی برای پشتیبانی از راهبرد بازخورد این است که اسناد مرتبط تشخیص داده شده توسط کاربر، حاوی اصطلاحاتی است که می‌تواند توصیف‌کننده خوشه بزرگتری از اسناد مرتبط باشد. چنین توصیفی از طریق تعامل نظام با کاربر حاصل می‌شود.

در روش خوشه‌بندی نتایج جستجو، مدارک بازیابی شده، به‌صورت خوشه‌های مرتبط و در مجاورت هم ارائه می‌شوند. بیشتر اوقات، پیمایش و نگاه کردن تعداد اندکی گروه همبسته، ساده‌تر از نگاه کردن به تک تک اسناد است. این روش، به‌ویژه در هنگامی که عبارت جستجو دارای جنبه‌ها یا معانی متفاوتی است، مفید است. عبارت "Clustering System" in "Information Retrieval" نمونه چنین جستجویی است که در موتور کاوش Clusty (ارائه شده توسط Vivisimo) انجام شده است (شکل ۳). همان‌طور که مشاهده می‌شود "نظام خوشه‌بندی" در "بازیابی اطلاعات" از جنبه‌های مختلفی قابل بحث است. تعداد کل نتایج، ۹۷ مورد بوده است که برخی از خوشه‌های اولیه آن در سمت چپ صفحه نمایش داده شده است، مانند Multimedia (7), Purpose (5), Web search (5), Conceptual... (8), Techniques (10)،

(5) Publications و چندین خوشه دیگر. شماره مقابل هر خوشه نشان‌دهنده تعداد نتایج مرتبط با هر خوشه است. ارائه نتایج، به‌صورت خوشه‌بندی می‌تواند رابط کاربر مؤثرتری برای درک آنچه در نتایج جستجو قرار دارد، نسبت به فهرست ساده و طولانی بدون ساختار نتایج در سایر موتور جستجوها، ارائه نماید (Manning, Raghavan, and Schutze 2008).

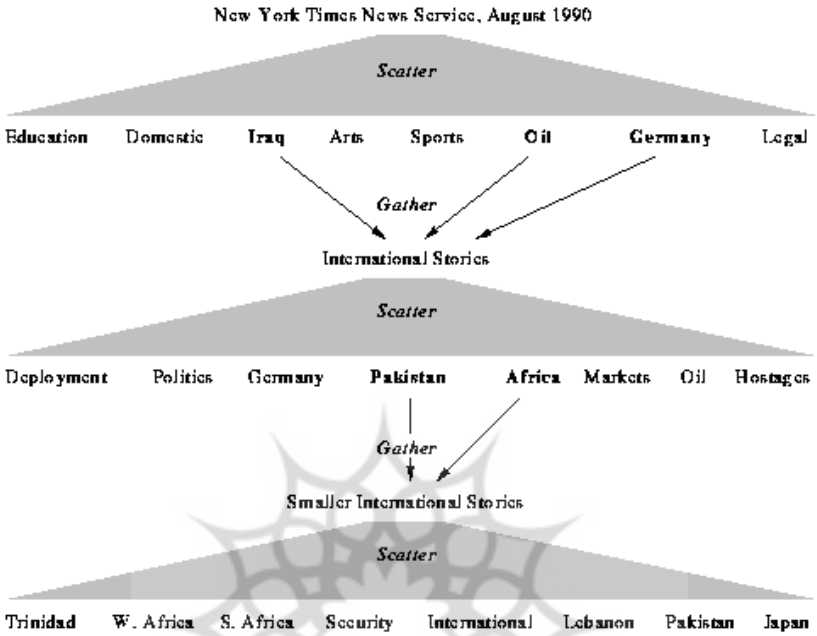


شکل۳. نمایش نتایج جستجوی عبارت
"Clustering System" in "Information Retrieval" در موتور جستجوی clusty

۵-۲. کاربرد دوم: پخش- جمع کردن (Scatter-gather)

هدف این روش ایجاد رابط کاربر بهتر است. در این روش، تمامی اسناد مجموعه در گروه‌هایی خوشه‌بندی می‌شود و کاربر می‌تواند از میان آن گروه‌ها انتخاب کند. گروه‌های انتخاب شده ادغام می‌شوند و مجموعه نتایج از نو خوشه‌بندی می‌گردد. این فرایند تا زمانی که خوشه‌ای از اسناد مورد علاقه به‌دست آید، تکرار می‌شود. نمونه این کاربرد در شکل ۴ نشان داده شده است.

اگر چه خوشه‌های تولید شده به این روش (به‌صورت خودکار)، به‌اندازه درخت موضوعی سلسله مراتبی راهنماهای تهیه شده توسط انسان دقیق نیستند و یافتن خودکار برچسب‌های توضیحی خوشه‌ها کاری پیچیده است، جستجوی مبتنی بر خوشه، بدلی جالب برای جستجوی کلیدواژه‌ای است که اکنون پارادایم بازیابی اطلاعات استاندارد است (Manning, Raghavan, and Schutze 2008). این حالت، به‌ویژه، هنگامی که کاربر در مورد انتخاب اصطلاح مورد استفاده مطمئن نیست و روش مرور را بر جستجوی کلیدواژه‌ای ترجیح می‌دهد، صادق است.



شکل ٤. نمونه‌ای از عملکرد خوشه‌بندی scatter-gather
(به نقل از Manning, Raghavan, and Schutze 2008)[1]

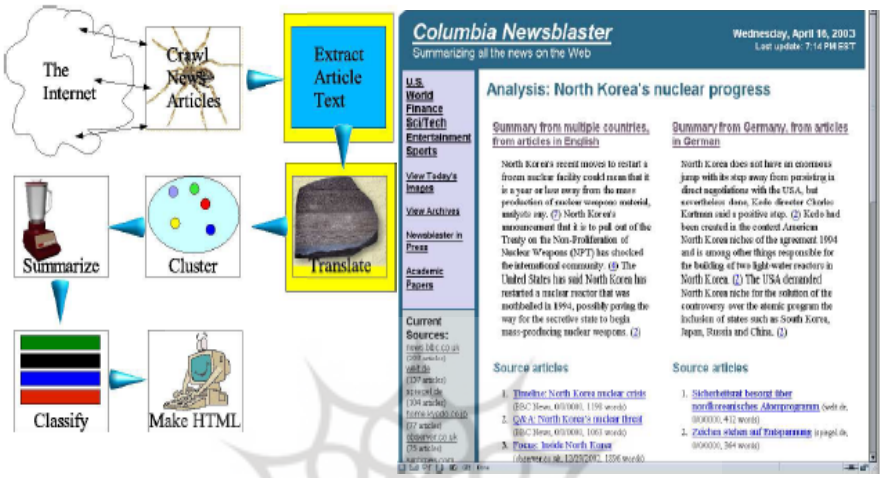
۵-۳. کاربرد سوم: خوشه‌بندی مجموعه

در این روش، به‌جای تکرار خوشه‌بندی توسط کاربر که در کاربرد پیشین، برای رسیدن به خوشه مطلوب صورت می‌گرفت، می‌توان یک خوشه‌بندی سلسله مراتبی را بر مبنای آمار، به‌صورت خودکار و بدون تعامل با کاربر، محاسبه کرد. Google news و پیشگام آن نظام Columbia Newsblaster (شکل ۵) نمونه‌هایی از این رویکرد هستند. در مورد اخبار، لازم است، به‌طور مکرر، خوشه‌بندی را محاسبه کرد تا مهمترین اخبار جاری در دسترس کاربران قرار گیرد. خوشه‌بندی برای دسترسی به مجموعه اخبار مناسب است، چون مطالعه اخبار بر مبنای

---

۱. مجموعه‌ای از داستان‌های خبری نیویورک تایمز در هشت خوشه در بالاترین سطح پخش شده است. کاربر سه خوشه از اینها را در مجموعه‌ای کوچک‌تر، یعنی (international stories) گردهم می‌آورد و یک بار دیگر عمل پخش اجرا می‌گردد. این فرایند تا زمانی که یک خوشه کوچک با اسناد مرتبط (Trinidad) یافت شود، ادامه دارد. جهت بررسی بیشتر چگونگی عملکرد scatter-gather به این نشانی مراجعه نمایید:
http://www2.parc.com/istl/projects/ia/sg-example1.html

جستجوی دقیق و واقعی صورت نمی‌گیرد، بلکه فرایندی برای انتخاب یک زیرمجموعه از داستان‌ها در مورد رویدادهای اخیر است.



شکل۵: سمت چپ، مراحل فرایند خوشه‌بندی
اخبار وب توسط Columbia Newsblaster و سمت راست، نمای رابط کاربر آن را نشان می‌دهد.

۵-٤. کاربرد چهارم: مدل‌سازی زبان

در این روش، به‌طور مستقیم از فرضیه خوشه‌بندی، بر اساس خوشه‌بندی تمام مجموعه، برای بهبود نتایج جستجو استفاده می‌شود. ابتدا، برای پاسخ پرسش، با استفاده از یک نمایه مقلوب استاندارد، مجموعه‌ای از اسناد اولیه ارائه می‌شود، اما بعد اسنادی از همان خوشه به سندهای بازیابی شده اضافه می‌شود، حتی اگر شباهت اندکی با پرسش داشته باشند. برای مثال، اگر پرسش راجع به "ماشین" است و چندین سند راجع به ماشین در خوشه اسناد مربوط به اتومبیل بازیابی شده است، می‌توان سایر اسناد همین خوشه را که اصطلاحاتی غیر از ماشین دارند (مانند خودرو یا اتومبیل) به آن اضافه نمود. این روش، می‌تواند بازیافت را افزایش دهد چون اسنادی که به‌صورت متقابل[1] با هم شباهت دارند، درکل، به هم ربط دارند. این ایده در مدل‌سازی زبانی مورد استفاده قرار گرفته است.

۵-۵. کاربرد پنجم: بازیابی مبتنی بر خوشه

همچنین، خوشه‌بندی تمام مجموعه می‌تواند سرعت جستجو را افزایش دهد. جستجو در مدل بردار فضایی نزدیک‌ترین اسنادی را که در مجاورت عبارت پرسش قرار دارند، پیدا می‌کند. در تنظیم سریع نزدیک‌ترین اسناد به پرسش در بازیابی اطلاعات استاندارد از نمایه مقلوب کمک

1. mutual

گرفته می‌شود، اما ممکن است که همواره امکان استفاده مؤثر از نمایه مقلوب وجود نداشته باشد مانند نمایه‌سازی معانی پنهان. در چنین مواردی، از محاسبه تشابه اسناد با پرسش می‌توان استفاده کرد، ولی این عمل کند است. فرضیه خوشه‌بندی، روش جایگزینی ارائه می‌کند: خوشه‌های نزدیک به پرسش را پیدا کنید و فقط اسناد درون این خوشه‌های نزدیک را مورد توجه قرار دهید. می‌توان درون این مجموعه‌های بسیار کوچک‌تر، تشابه و رتبه‌بندی آنها را به‌روش معمول بسیار دقیق و کامل انجام داد. از آنجا که تعداد خوشه‌های به‌دست آمده از تعداد کل اسناد مجموعه کم‌تر است، یافتن نزدیک‌ترین خوشه بسیار سریع انجام می‌گیرد و به‌دلیل اینکه اسنادی که با پرسش تطابق دارند، خود باهم مشابه هستند در یک خوشه ظاهر می‌شوند. اگر چه این الگوریتم دقیق نیست، کاهش کیفیت ناچیز است (Manning, Raghavan, and Schutze 2008).

٦. ارزیابی خوشه‌بندی

طبق گفته یو و هو، نظام‌های خوشه‌بندی به سه روش مورد ارزیابی قرار گرفته‌اند (Yoo and Hu 2006). تقسیم‌بندی آنان را می‌توان به‌صورت جدول ۲ نمایش داد. گفته زو و وانش نیز بیانگر این است که به‌طور کلی، سه دسته از معیارهای ارزیابی خوشه‌بندی وجود دارد که عبارتند از معیارهای بیرونی و معیارهای درونی و معیارهای نسبی (Xu and Wunsch II 2005).

جدول ۲. روش‌های ارزیابی نظام خوشه‌بندی

روش ارزیابی	هدف	تکنیک	کاربرد	نمونه‌ها
کاربرمدار (رضایت کاربر)	سنجش اثربخشی نتایج موتور کاوش خوشه‌بندی	سنجش رضایت کاربر از خروجی نظام	جامعه بازیابی اطلاعات که مطالعات هدف‌مدار انجام می دهند.	استفاده از فرم نظر سنجی یا پرسشنامه
عملکردهای واقعی	ارزیابی الگوریتم‌های خوشه‌بندی	مقایسه دقت و بازیافت الگوریتم‌های خوشه‌بندی	در خوشه‌بندی مسطح کاربرد دارد.	K-means[1]

←

۱ . میانگین‌های K، مهم‌ترین الگوریتم خوشه‌بندی مسطح است و هدف آن به حداقل رساندن مجذور مربع فاصله اقلیدسی اسناد از مراکز (کانون‌های) خوشه‌هایشان است.

ادامهٔ جدول ۲. روش‌های ارزیابی نظام خوشه‌بندی

روش ارزیابی	هدف	تکنیک	کاربرد	نمونه‌ها
استانداردهای طلایی یا کلیدهای پاسخ	*سنجش میزان تطابق الگوریتم خوشه‌بندی با استانداردهای حرفه‌ای*	*مقایسه خروجی الگوریتم‌ها با مجموعه پاسخ‌های کلیدی یا استانداردهای طلایی که توسط متخصصان تأیید شده است.*	*پژوهش‌های آزمایشگاهی نظام‌های خوشه‌بندی*	*Mutual information metrics, Misclassification index, Purity, Confusion matrix, F-measure, Entropy*

مانینگ، راگهاون و شوتز در مورد سه رویکرد ارزیابی خوشه‌بندی چنین توضیح می‌دهند:

«به‌طور معمول، کارکردهای عینی[1] خوشه‌بندی، تعیین‌کننده هدف دستیابی به شباهت زیاد اسناد درون هر خوشه و حداقل شباهت بین خوشه‌های مجزاست. میزان شباهت اسناد درون خوشه با هم و تمایز بین خوشه‌ها معیاری درونی[2] برای کیفیت خوشه‌بندی است، اما به‌دست آوردن نمره به‌نسبت بالا، بر اساس معیار درونی به‌معنی اثربخشی در کاربرد نیست. روش دیگر، ارزیابی مستقیم در هنگام کاربرد بر اساس علاقه، به‌جای معیار درونی است. برای خوشه‌بندی نتایج جستجو، ممکن است زمانی که کاربر برای دستیابی به پاسخ از طریق الگوریتم‌های خوشه‌بندی متفاوت صرف می‌کند، اندازه گرفته شود. این روش، کامل‌ترین ارزیابی است، اما گران تمام می‌شود، به‌ویژه هنگامی که تعداد زیادی از کاربران در مطالعه شرکت نمایند. به‌جای استفاده از قضاوت کاربر، می‌توان از مجموعه خوشه‌های ارزیابی شده بر اساس فاصله‌ها یا استاندارد طلایی که توسط متخصصان مورد توافق است، استفاده کرد. استاندارد طلایی، مجموعه‌ای از خوشه‌های ایده‌آل است که مبتنی بر قضاوت و توافق متخصصان، در مورد تعداد خوشه‌های ممکن برای هر مقوله، ایجاد شده است. می‌توان از این معیار بیرونی برای مقایسه میزان قابلیت نظام خوشه‌بندی در تفکیک خوشه‌های از پیش تعریف شده که به استاندارد طلایی مشهور است، استفاده نمود» (Manning, Raghavan, and Schutze 2008).

معیارهای سنجش کیفیت خوشه‌بندی (Yoo and ; Manning, Raghavan, and Schutze 2008 Hu 2006) عبارتند از:

1. objective functions 2. internal criterion

۱. **خلوص**[1]، یک معیار ارزیابی بیرونی و بیانگر درصد اسناد مربوط به موضوع غالب در یک خوشه است. برای محاسبه خلوص، میانگین وزن‌ها را که از طریق محاسبه درصد دقت به‌دست می‌آید، اندازه می‌گیرند. هر خوشه در طبقه‌ای قرار داده می‌شود که بیشترین تکرار را در آن خوشه دارد، سپس صحت این تخصیص از طریق شمارش تعداد اسنادی که درست اختصاص داده شده‌اند و تقسیم بر N (تعداد کل اسناد) محاسبه می‌گردد. خوشه‌بندی نامناسب دارای ارزش خلوص پایین نزدیک به صفر و خوشه‌بندی کامل دارای درجه خلوص یک است. هنگامی که تعداد خوشه‌ها زیاد باشد، به‌دست آوردن درجه خلوص بالا کار ساده‌ای است، به‌ویژه اگر هر سند دارای خوشه متعلق به خودش باشد، خلوص برابر یک خواهد بود. بنابراین، نمی‌توان خلوص را به‌عنوان کیفیت خوشه‌بندی در مقابل تعداد خوشه‌هایی که انتخاب می‌شوند، پذیرفت.

۲. **اطلاعات تعاملی نرمال شده**[2]، که رویکرد تفسیر نظری–اطلاعاتی[3] است. این روش، بر مبنای احتمالات، آنتروپی، تخمین، و بسامد نسبی صورت می‌گیرد. به‌دلیل نرمال بودن، می‌توان از آن برای مقایسه خوشه‌بندی‌های دارای تعداد خوشه‌های متفاوت استفاده نمود. آنتروپی[4]، بیانگر چگونگی توزیع طبقه‌های K در هر خوشه است. برای محاسبه آن، همانند خلوص و مقیاس F[5]، از محاسبه میانگین وزن استفاده می‌شود.

۳. **شاخص راند**[6]، که به خوشه‌بندی به‌عنوان سلسله‌ای از تصمیمات، نگاه می‌کند و درصد خطاهای مثبت و منفی تصمیم‌گیری در طول خوشه‌بندی را مشخص می‌کند. شاخص اشتباه طبقه‌بندی[7]، نسبت تعداد اشیاء اشتباه طبقه‌بندی شده به کل مجموعه است.

۴. **مقیاس F**، اشکال شاخص راند این است که برای تصمیمات مثبت و منفی به یک اندازه ارزش قائل است و وزن مساوی در نظر می‌گیرد. اما مقیاس F، علاوه بر آن دو نوع خطا، از تمایز وزن‌دهی نیز پشتیبانی نموده و به بازیافت، وزن بیشتری اختصاص می‌دهد. مقیاس F دقت و بازیافت حاصل از بازیابی اطلاعات را با هم ترکیب می‌کند. مزیت این مقیاس این است که جامعه پژوهشی از قبل با آنها آشناست.

هر چه مقدار شاخص اشتباه طبقه‌بندی و میزان آنتروپی کوچک‌تر باشد، کیفیت خوشه‌بندی بهتر است. اما، هر چه مقیاس F و خلوص بزرگتر باشد، بیانگر خوشه‌بندی باکیفیت‌تری است. به‌عبارت دیگر، تمایز[8] و تراکم[9] دو سنجه مفید برای ارزیابی خروجی الگوریتم خوشه‌بندی هستند. پژوهشگران به‌دنبال تراکم اندک و تمایز زیاد هستند (Harper, Sen, and Frankowski 2007).

---

1. purity
2. normalized mutual information
3. information-theoretical interpretation
4. entropy
5. F measure
6. Rand index
7. misclassification index
8. separation
9. compactness

۷. نقطه تماس نظام‌های خوشه‌بندی با کاربران

مطالب نظری بیان شده فوق مبنای طراحی نظام‌های خوشه‌بندی جهت بازیابی اطلاعات مرتبط با نیاز کاربران است. نقطه تماس نظام با نیازمندان اطلاعات، رابط کاربر است. هر موتور خوشه‌بندی سعی دارد در رابط کاربر خود عناصری قرار دهد که به بهترین وجه ویژگی‌هایش را به‌نمایش بگذارد تا کاربر به‌راحتی بتواند با نظام تعامل داشته باشد. از طرف دیگر، فرض بر این است که به‌کارگیری برخی عناصر در تعداد زیادی از رابط کاربر موتورهای خوشه‌بندی می‌تواند دلیلی بر اهمیت به‌کارگیری آن عنصر در چنین نظام‌هایی باشد. بخش بعدی این نوشته در پی شناسایی عناصر به‌کار گرفته شده در اکثر موتورهای خوشه‌بندی مورد بررسی است. شناسایی چنین عناصری می‌تواند توجه طراحان موتورهای خوشه‌بندی فارسی‌زبان را به خود جلب نماید (شواهد نشان می‌دهد که تاکنون موتور خوشه‌بندی برای منابع فارسی عرضه نشده است).

۸. پیشینه

استفاده از خوشه‌بندی در رشته‌های مختلف مانند زیست‌شناسی، روان‌پزشکی، روان‌شناسی، باستان‌شناسی، زمین‌شناسی، جغرافیا، و بازاریابی سابقه طولانی دارد. همچنین، علاقه به استفاده از آن در زمینه شناسایی الگوها و پردازش تصاویر پیوسته رو به افزایش بوده است (Jain, Murty, and Flynn 1999). مطرح نمودن کاربرد خوشه‌بندی در حوزه بازیابی اطلاعات به سالتون (Salton 1991) برمی‌گردد. اما، در زمینه رابط کاربر خوشه‌بندی مقاله زمیر و اتزیونی به‌خاطر توجه به برخی معیارها و عناصر رابط کاربر در موتور جستجوی خوشه‌بندی Grouper، جزء موارد پر استناد در این حوزه است. این دو، به احتمال، اولین کسانی هستند که در مورد عناصر و ویژگی‌های رابط کاربر موتور خوشه‌بندی بحث نموده‌اند(Weiss 2001). از دیدگاه آنها، ملزومات نظام خوشه‌بندی نتایج بازیابی اسناد عبارتند از: ربط، خلاصه‌های قابل مرور، سرعت، بافت سازگار[1]، قابلیت افزایش[2]، همپوشانی خوشه‌ها، تعاریف ساده خوشه‌ها، و استفاده از عبارات (Zamir and Etzioni 1999).

همچنین، کتاب هیرست با عنوان *Search User Interfaces*، از جمله کتاب‌های ارزشمند در مورد جنبه‌های مختلف رابط کاربر، به‌طور کلی، است. اگر چه هیرست، به‌طور خاص، در مورد رابط کابر موتورهای خوشه‌بندی صحبت نکرده است، در مورد کاربرد خوشه‌بندی برای تقویت نمایش نتایج جستجو و کاربرد مصورسازی در رابط کاربر و تکنیک‌های آن در فصل‌های ۸ و ۱۰ کتابش مطالب ارزشمندی ارائه نموده است (Hearst 2009).

اگر چه نزدیک به دو دهه است که استفاده از خوشه‌بندی در بازیابی اطلاعات پیشنهاد شده است، کاربرد خوشه‌بندی در بازیابی اطلاعات در میان جامعه کتابداری و اطلاع‌رسانی و طراحان

1. snippet tolerance 2. incrementality

موتورهای جستجوی زبان فارسی، به‌نسبت موضوعات دیگر، کمتر مورد توجه قرار گرفته است. به‌همین دلیل، در زمینه رابط کاربر نظام‌های خوشه‌بندی، به‌طور خاص، تاکنون کار نشده است، اما در زمینه معیارهای کلی رابط کاربر، رهنمودها (چنگیزی ۱۳۸۷؛ زره‌ساز و فتاحی ۱۳۸۵) و کارهای تجربی ارزشمندی (نوکاریزی، فتاحی، و داورپناه ۱۳۸۶؛ اسفندیاری مقدم و بهاری موفق ۱۳۸۸) منتشر شده است.

۹. اهمیت توجه به رابط کاربر در نظام‌های خوشه‌بندی

راهبردهای بازخورد ربط از طریق خوشه‌بندی را می‌توان به‌طور مستقیم، در رابط کاربرهای گرافیکی نظام‌های اطلاعاتی نوین به‌کار برد. در هر صورت، از آنجا که اکنون تعامل نظام با کاربر دارای اهمیت بیشتری است (حریری ۱۳۸۳)، فنون کسب اطلاعات بازخورد از کاربر درخور توجه است. هدف، فراهم کردن امکان تعیین روابط اسناد ارائه شده (از طریق ابزارهای بصری) برای پاسخ، توسط کاربر است.

ترکیب خوشه‌بندی محلی، تحلیل سراسری، ارائه نتایج به‌صورت بصری، و رابط کاربرهای تعاملی از مسائل پژوهشی مهم به‌شمار می‌آیند. فراهم کردن امکان جستجو و بررسی مصور فضای اسناد برای کاربر و ارائه کلیدهایی که بتواند همراه فرایند فرمول‌بندی پرسش مورد استفاده قرار دهد، از مباحث مهم ربط است. نتایج مثبت این حوزه پژوهشی ممکن است سبب توجه به طراحی رابط کاربرهای نظام‌های خوشه‌بندی شود و امکان دارد به‌طور وسیع، مورد توجه محققان قرار گیرد (Baeza-Yates and Riberio-Neto 1999).

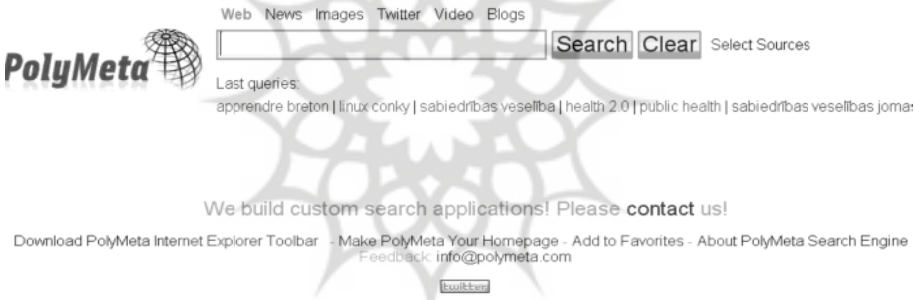
به‌طور کلی، اصلاح پرسش کاربر از دو طریق صورت می گیرد: استفاده از بازخورد ربط و استفاده از اصطلاحنامه کمکی (Baeza-Yates and Riberio-Neto 1999; Cunningham et al. 1998). برای استفاده از بازخورد ربط نظام‌های خوشه‌بندی در وب، چگونگی طراحی و عناصر قرار داده شده در محیط رابط کاربر از اهمیت زیادی برخوردار است. هر چه عناصر قرار داده شده در رابط کاربری نظام‌های خوشه‌بندی ساده، مرتبط و دقیق باشد، کاربر بهتر می‌تواند متناسب با نیاز اطلاعاتی خود با سیستم تعامل داشته و پرسش خود را اصلاح نماید.

از طرف دیگر، پیش فرض لازم جهت بهبود عملکرد بسیاری از الگوریتم‌های خوشه‌بندی و دستیابی به مناسب‌ترین خوشه، بر مبنای بازآرایی خوشه‌ها بر اساس انتخاب‌های کاربر و تکرار فرایند خوشه‌بندی است (Xu and Wunsch II 2005). به این دلایل، لازم است که رابط کاربرهای خوشه‌بندی دارای عناصری باشند که ویژگی‌هایی چون تعاملی، کاربرپسندی، انعطاف‌پذیری، راهنمایی، و سادگی را فراهم نمایند.

از آنجا که موتور خوشه‌بندی در زبان فارسی موجود نیست، مقاله حاضر عناصر مشترک و مهم موجود در رابط کاربر ۱۰ موتور خوشه‌بندی وب را که در متون مربوط به کاربرد خوشه‌بندی در بازیابی اطلاعات، مورد مطالعه قرار گرفته‌اند، بررسی کرده است.

۱۰. عناصر و مؤلفه‌های موجود در رابط کاربر موتورهای خوشه‌بندی

در این قسمت، ابتدا با استفاده از تصویر، اجزاء مربوط به رابط کاربر یکی از موتورهای خوشه‌بندی و کارکرد عناصر و مؤلفه‌های رابط کاربر موتور خوشه‌بندی توضیح داده شده است. در پایان، جدول وضعیت استفاده موتورهای خوشه‌بندی مورد مطالعه از عناصر شناسایی شده، آمده است. از آنجا که به‌نظر می‌رسد موتور خوشه‌بندی PolyMeta[1]، از نظر تقسیم‌بندی و تعداد مؤلفه‌های رابط کاربر مفصل‌تر از بقیه باشد، بیشتر تصاویر، نمونه برشی از آن است. در صورت وجود عناصر بیشتر در سایر موتورهای خوشه‌بندی مورد مطالعه، تصاویر آنها نیز ارائه و توضیح داده شده است.



شکل ٦. نمونه سرصفحه و پاصفحه موتور خوشه‌بندی

- صفحه اول رابط کاربر (شکل ۶)

۱. منوی جستجو که دارای گزینه‌های جستجو و پاک کردن عبارت جستجو است.

۲. قالب/نوع منابعی که موتور خوشه‌بندی قادر به جمع‌آوری اطلاعات از آنهاست، مانند: وب، تصاویر، اخبار، وبلاگ‌ها، ویدیوها، و غیره.

۳. موتور کاوش‌هایی که موتور خوشه‌بندی از نمایه آنها استفاده می‌نماید. این موتورها از طریق پیوند Select Sources قابل مشاهده است (شکل ۷). کاربر با انتخاب موتورهای

---

۱. PolyMeta: یک ابرموتور جستجو است که کار خوشه‌بندی را انجام می‌دهد و توسط تولیدکننده نظام‌های مجارستان در K-Prog ایجاد شد. این نظام، موتور پشت پرده خدمات دسترسی آزاد TOXSEEK است و بیش از ۵۰ پایگاه اطلاعاتی دولتی را که به مواد سمی و موضوعات مرتبط می‌پردازند جستجو می‌کند (Jacso 2007).

بیشتر، می‌تواند دامنه جستجوی منابع موجود در وب را از طریق نمایه موتورهای کاوش موجود در فهرست، بسط دهد.

۴. در پایین صفحه، مسائل حقوقی (حقوق مالکیت فکری) و امکانات دیگری مانند استفاده از راهنما، مطالعه توضیحات مربوط به موتور خوشه‌بندی حاضر، پیوند قرار دادن موتور خوشه‌بندی به‌عنوان صفحه اول جستجوگر وب، امکان بارگذاری منوی جستجوگر اینترنت ویژه موتور خوشه‌بندی، امکان تماس، و بازخورد نظرات کاربران وجود دارد.

به‌طور کلی، آنچه از صفحه اول برداشت می‌شود سادگی/عدم شلوغی صفحه و خوانایی و انعطاف‌پذیری است، به‌طوری که کاربر، به‌راحتی می‌تواند به‌دلخواه، نوع و گستره منابع مورد نظر را برای جستجو، تغییر دهد.

شکل ۷. موتورهای جستجویی که در پیوند Select Sources کاربر می‌تواند انتخاب نماید تا نمایه آنها توسط موتور خوشه‌بندی برای جستجوی مورد نظر مورد استفاده قرار گیرد.

- پس از انجام جستجوی اولیه، بخش‌های زیر نمایش داده می‌شود (شکل ۸)

۱. Search within به قسمت منوی جستجو اضافه می‌شود. با انتخاب این گزینه، عبارت جستجو در خوشه‌ها و منابع نمایش داده شده پررنگ‌تر می‌گردد.

۲. گزینه بعدی Show/Hide hits است که کاربر می‌تواند با کلیک روی آن، فهرست موتورهای انتخاب شده در Select Sources را که عبارت جستجو در نمایه آنها یافت شده و نیز آمار تعداد عبارت‌ها را نمایش دهد یا مخفی نماید.

۳. در زیر منوی جستجو، عبارت Also consider اضافه می‌شود و عبارت‌های پیشنهادی مرتبط با عبارت جستجو نمایش داده می‌شود.

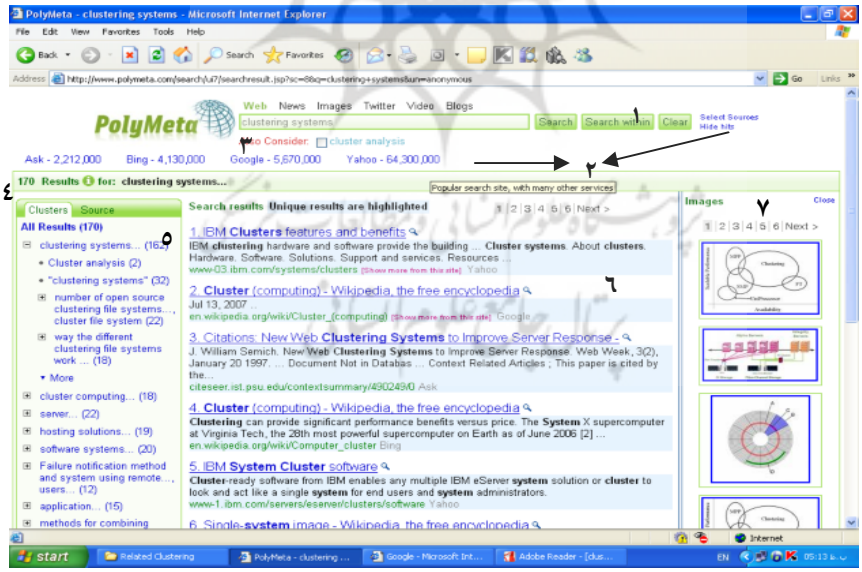
۴. در قسمت Results for، عبارت جستجو و تعداد نتایج بازیابی شده آورده می‌شود. با قرار دادن ماوس روی حرف " i " برچسب توضیحی ظاهر می‌گردد که بیانگر نحوه عملکرد موتور خوشه‌بندی است.

۵. در قسمت ۵ (منوی عمودی سمت چپ شکل ۸) دو منوی Clusters و Sources و محتویات آنها، به‌صورت همپوشان، ظاهر می‌شود. در برگه Clusters، تمامی خوشه‌های کلی و زیرشاخه‌های موضوعی آنها با تعداد اسناد قرار گرفته در هر خوشه یا زیرشاخه

نمایش داده می‌شود (علامت " + " به‌معنی وجود زیرشاخه در خوشه مربوط است، علامت " – " به‌معنی نمایش زیرخوشه‌هاست و علامت " ▪ " به‌معنی نمایش منبع بازیابی شده است). همچنین، در برگه Sources فهرست موتورهایی که نمایه آنها مورد جستجو قرار گرفته و عبارت جستجو در آنها یافت شده است با تعداد اسناد یافت شده در مقابل هرکدام نمایش داده می‌شود.

۶. بخش وسیعی از مرکز صفحه به فهرست نتایج جستجو همراه با جزئیات آنها اختصاص دارد (قسمت ۶ شکل ۸). محتویات این صفحه با انتخاب خوشه‌های اصلی و زیرشاخه‌های آنها در منوی clusters توسط کاربر عوض شده و اسناد مرتبط با موضوع خوشه انتخاب شده، نمایش داده می‌شود. اطلاعات ارائه شده در این قسمت، دارای تنوع و انعطاف و قابلیت‌های مختلف است.

۷. در ستون سمت راست شکل ۸ منوی Images که حاوی تصاویر مربوط به عبارت جستجو است، نمایش داده شده است. در بالای این منو، دو گزینه Next و Close تعبیه شده است که کاربر می‌تواند به تورق این ستون بپردازد و تصاویر بیشتری را که در صفحات بعدی قرار دارند، بیابد یا اینکه این منو را ببندد و از کل فضای صفحه برای نمایش محتویات سایر منوها استفاده نماید.



شکل ۸. قسمت‌های نمایش داده شده بعد از وارد کردن عبارت جستجو و اجرای آن توسط کاربر (قسمت‌های افزوده شده با شماره ۱ تا ۷ مشخص شده است.)

به‌طور خلاصه، محتویات صفحه نمایش نتایج، ویژگی‌های انعطاف‌پذیری، سادگی، خوانایی، کاربرپسندی، گزینه‌های اختیاری، کمک، و توضیح در هنگام نیاز را داراست و از تصویر و متن برای توسعه دید کاربر بهره گرفته است. اما، عناصر دیگر که در رابط کاربر برخی موتورهای خوشه‌بندی (به‌جز PolyMeta) وجود دارد، به‌شرح زیر است:

۱. جستجوی پیشرفته: نمونه آن در شکل ۹ که برشی از موتور خوشه‌بندی Clusty است، نشان داده شده است. کاربر، از طریق گزینه‌های آن می‌تواند جستجوی دقیق‌تری انجام دهد.

شکل ۹. منوی جستجوی پیشرفته موتور خوشه‌بندی Clusty که شامل گزینه‌های متعدد برای محدود کردن جستجو براساس علایق کاربر است.

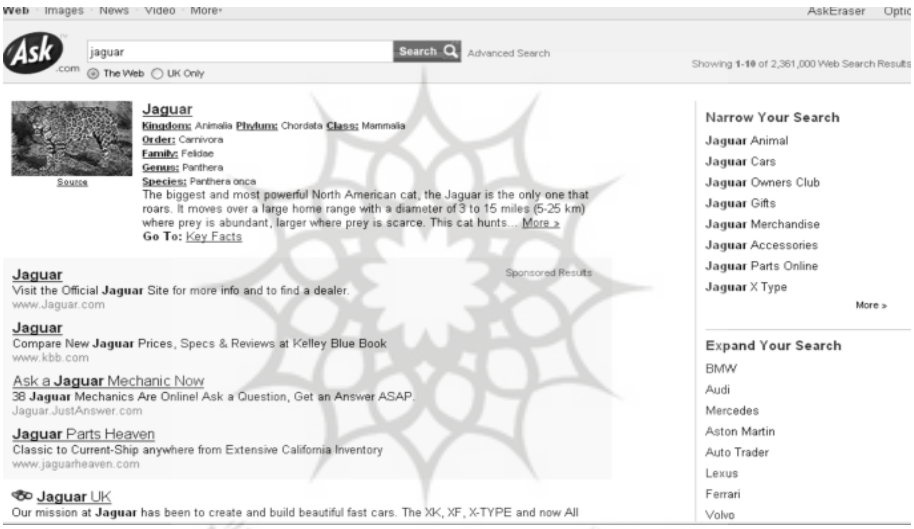
۲. گزینه دیگر موتور خوشه‌بندی Clusty مربوط به تنظیمات/ ترجیحات (Preferences/settings) است که از طریق آن، کاربر می‌تواند، به‌دلخواه، تعداد نتایجی که در صفحه نمایش داده می‌شود، تعداد خوشه‌هایی که نمایش داده خواهد شد، مکان نمایش نتایج (پنجره کنونی یا پنجره‌ای جدید)، نوع گزینه‌ها و متعلقات مربوط به نتایج که باید نشان داده شود، فیلتر بزرگسالان، کشور، زبان، و چند گزینه دیگر را تنظیم نماید.

۳. در منوی سمت چپ موتور کاوش Clusty علاوه بر زبانه منوی Clusters و Sources، زبانه Sites تعبیه شده است که در آن، نوع سایت‌هایی که منابع از آنها بازیابی شده است و تعداد اسنادی که از هر نوع یافت شده است، نمایش داده می‌شود. در سمت چپ هر کدام علامت "+" قرار گرفته باشد، به این معنی که دارای زیر شاخه‌هایی است و می‌توان با کلیک روی آن زیرشاخه‌های آن را نمایش داد.

۴. موتور خوشه‌بندی Ask دارای چند عنصر مفید دیگر است (شکل ۱۰). در سمت راست، دو منوی مربوط به گسترش و محدود کردن عبارت جستجو مشاهده می‌شود که در منوی

گسترش، عبارت‌های کلی‌تر (اعم) از عبارت جستجو و در منوی محدود کردن، عبارت‌های جزئی‌تر (اخص)- به‌طور عمده، ترکیب عبارت جستجو با عبارت‌های دیگر است- به کاربر پیشنهاد می‌شود. این عمل کاری شبیه اصطلاحنامه مرتبط با عبارت جستجو است.

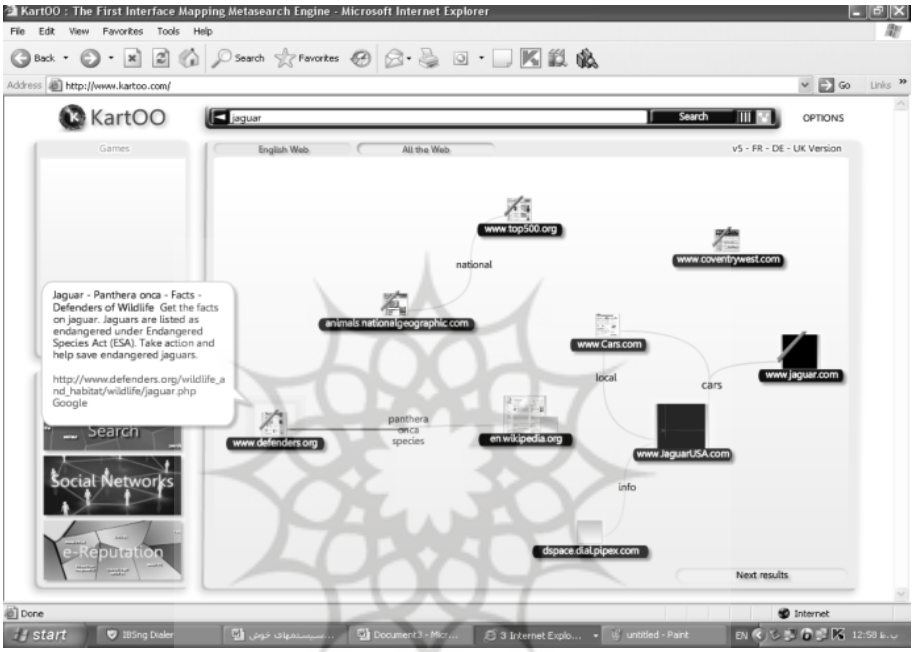
۵. نکته دیگر اینکه کاربر می‌تواند جستجوی خود را در تمام وب انجام دهد یا محدود به کشور انگلستان نماید (شکل ۱۰ زیرمنوی جستجو). گزینه محدود کردن به کشور یا ناحیه‌ای ویژه، ممکن است، در مورد زبان‌های محدود، برای برخی کاربران مفید باشد.



شکل ۱۰. موتور خوشه‌بندی Ask با دو منوی گسترش جستجو و محدود کردن جستجو (سمت راست تصویر) و محدود کردن دامنه جستجو به کشور انگلستان UK (در زیرمنوی جستجو)

۶. عنصر دیگری که از ویژگی‌های مهم موتورهای خوشه‌بندی به‌شمار می‌رود، گزینه‌های مربوط به نمایش متنی و گرافیکی نتایج بازیابی است. این عناصر در موتور خوشه‌بندی Kartoo وجود دارد (شکل ۱۱). کاربر با کلیک روی گزینه متن (در انتهای سمت راست منوی جستجو)، می‌تواند نتایج بازیابی را به‌صورت متن، نشان دهد. با انتخاب گزینه نمایش گرافیکی، می‌تواند روابط نتایج بازیابی شده را به‌صورت شبکه‌ای از وبلاگ‌ها و وب‌سایت‌های مرتبط به هم که با خطوط به یکدیگر وصل شده‌اند، مشاهده نماید. در این قسمت، با قرار دادن ماوس روی هر گره از این شبکه، منویی متنی که توضیحی در مورد عبارت جستجو و منبع بازیابی شده است، نمایش داده می‌شود.

۷. گزینه دیگر، قابلیت نمایش سابقه جستجو است که در انتهای سمت چپ منوی جستجو قرار دارد. کاربر با کلیک روی این گزینه، می‌تواند عبارت‌های جستجویی که پیش از این، برای جستجو به‌کار برده است، مشاهده نماید (شکل ۱۱).



شکل ۱۱. موتور خوشه‌بندی Kartoo

۱۱. کاربرد عناصر و مؤلفه‌ها در موتورهای خوشه‌بندی

وضعیت استفاده از عناصر و مؤلفه‌های رابط کاربر (شناسایی شده در ۱۰ موتور خوشه‌بندی مورد مطالعه) در جدول ۳ نشان داده شده است. علامت " * " نشانه وجود و علامت " – " نشانه عدم وجود عنصر در رابط کاربر است. برخی از این عناصر مانند جستجوی ساده، جستجوی پیشرفته، تنظیمات/ترجیحات، قالب/نوع منابع، امکان بازخورد و اطلاعات تماس، منوهای توضیحی، فهرست خوشه‌های اصلی و فرعی، فهرست نتایج جستجو (شامل عنوان، بافت عبارت جستجو، URL منبع اصلی، موتورهایی که منبع اصلی را نمایه کرده‌اند، امکان تورق نتایج، و نمایش در صفحات پشت سرهم)، نمایش متنی و تصویری (نه گرافیکی)، و فیلتر بزرگسالان، دست کم، در نصف رابط کاربرهای مورد مطالعه وجود دارد.

جدول ۳: وضعیت استفاده از عناصر و مؤلفه‌ها در رابط کاربر موتورهای خوشه‌بندی مورد مطالعه

موتورهای خوشه‌بندی عناصر	PolyMeta	Clusty	Kartoo	Ask	Carrot2	WebClust	Snake T	iBoogie ™	Xsearch	BizNetic
جستجوی ساده (Simple search)	*	*	*	*	*	*	*	*	*	*
جستجوی پیشرفته (Advanced search)	–	*	–	*	*	–	–	*	–	*
تنظیمات/ترجیحات (references/ setting)	–	*	–	*	*	–	*	*	–	*
قالب/نوع منابع (format/types of resources)	*	*	–	*	*	–	–	*	*	*
انتخاب منابع (Select sources)	*	–	–	–	–	–	–	*	–	–
اطلاعات تماس (About/feedback/contacts)	*	*	–	*	*	*	*	*	*	*
جستجو در نتایج بازیابی شده (Search within)	*	–	–	–	–	–	–	–	*	–
نمایش/مخفی کردن برخوردها (Show/hide hits)	*	–	–	–	–	–	–	–	–	–
موارد مشابه را نیز جستجو کنید (Also consider/find similar)	*	–	–	–	–	–	–	–	–	*
منوهای کمک (Pop up menus)	*	–	–	–	*	–	–	*	*	*
خوشه‌ها (Clusters)	*	*	–	–	*	*	*	*	*	*
منابع (Sources)	*	*	–	–	–	–	–	*	–	–
دامنه سایت‌ها (Sites domain)	–	*	–	–	*	–	–	–	–	*
نتایج جستجو (Search results)	*	*	*	*	*	*	*	*	*	*
عنوان(Title)	*	*	*	*	*	*	*	*	*	*
بافت کاربرد عبارت (Context of term usage)	*	*	*	*	*	*	*	*	*	*
یو.آر.ال. اصلی مقاله (URL of original doc.)	*	*	*	*	*	*	–	*	*	*
موتور جستجو(ها)ی نمایه‌کننده (Fetcher search engine(s))	*	*	*	–	*	–	*	*	–	–

←

ادامهٔ جدول ۳: وضعیت استفاده از عناصر و مؤلفه‌ها در رابط کاربر موتورهای خوشه‌بندی مورد مطالعه

موتورهای خوشه‌بندی عناصر	PolyMeta	Clusty	Kartoo	Ask	Carrot2	WebClust	Snake T	iBoogie™	Xsearch	BizNetic
پیش‌نمایش سند (Doc. Preview)	*	*	–	*	*	–	–	–	–	–
نمایش دادن/پنهان کردن نتایج بیشتر از این سایت (Show/hide more hits from this site)	*	–	–	–	–	–	–	–	–	–
بعدی/قبلی (Next/Previous)	*	*	*	–	–	*	–	*	–	*
پیشنهاد عبارت‌های اعم/اخص (Suggest Narrow/Broad terms)	–	–	–	*	–	–	–	–	–	–
محدود کردن جستجو به زبان/کشور Limit search to country/language	–	*	–	*	*	–	–	–	–	*
نمایش گرافیکی/تصویری (Graphical/visualization view)	–	*	*	–	*	–	–	–	–	–
نمایش متن (Text view)	*	*	*	*	*	*	*	*	*	*
سابقه عبارت جستجو (Search term history)	–	–	*	–	–	–	–	–	–	–
فیلتر بزرگسالان (Adult filter)	–	*	*	–	*	–	–	*	–	*
نسخه زبان‌های مختلف (En Fr Ger D version)	–	–	*	–	–	–	–	–	–	–
جمع عناصر	۱۹	۱۹	۱۲	۱۳	۱۸	۹	۹	۱۷	۱۱	۱۷

برخی ویژگی‌ها نیز منحصر به یک یا چند رابط کاربر مورد مطالعه است، مانند نسخه‌های متعدد برای کشورهای مختلف یا امکان نمایش گرافیکی خوشه‌ها و ارتباط آنها. نکته قابل توجه در مورد کاربرد این عناصر در رابط کاربرهای مورد مطالعه این است که اگر چه ممکن است یک عنصر در چند رابط کاربر استفاده شده باشد، سطح، تعدد، عمق، یا حوزه جغرافیایی آنها با هم برابر نیست و میزان منابع، انواع/قالب یا موتورهای جستجویی که نمایه آنها مورد استفاده موتور خوشه‌بندی است، با هم فرق دارد. برخی از آنها از یک موتور جستجو (مانند استفاده موتور خوشه‌بندی carrot2 از موتور جستجوی !Yahoo) بیشتر از دیگر موتورهای جستجوی موجود در وب بهره می‌گیرند و یا راجع به یک کشور خاص جستجوی ویژه دارند (مانند جستجوی منابع مربوط به UK در موتور خوشه‌بندی Ask).

از نظر تعداد استفاده از عناصر، PolyMeta و Clusty (با ۱۹ عنصر)، Carrot2 (با ۱۸ عنصر)، و iBoogie™ وBizNetic (با ۱۷ عنصر) در رده اول تا سوم قرار دارند. اگر چه Kartoo از تعداد عناصر محدودی بهره می‌گیرد، نمایش گرافیکی قوی و منحصر به‌فردی دارد.

یادآور می‌شود که برخی از عناصر شناسایی شده، منحصر به موتورهای خوشه‌بندی است و برخی در موتورهای کاوش پیشین هم سال‌ها به‌کار رفته است. دلیل این امر، همان‌طور که پیشتر اشاره شد، این است که موتورهای خوشه‌بندی مکمل جستجوی کلیدواژه‌ای و مروری هستند نه جانشینی برای آنها.

۱۲. بحث پایانی

کاربردهای خوشه‌بندی در بازیابی اطلاعات مدتهاست که مورد توجه نظام‌های خودکار قرار گرفته است. این کاربردها شامل خوشه‌بندی زبانی و مجموعه اسناد و نتایج بازیابی است. با توجه به عملکردهای عینی خوشه‌بندی، در هر نظام از الگوریتم‌های خوشه‌بندی متفاوتی استفاده می‌شود. الگوریتم‌های خوشه‌بندی، هر کدام بر اساس انتخاب یا استخراج ویژگی‌هایی از اسناد عمل می‌نمایند تا کارکرد عینی آنها را برآورده کنند. فرضیه خوشه‌بندی، بر اساس پاسخگویی یا رفتار مشابه اسناد یک خوشه در برابر پرسش کاربر است، به عبارت دیگر، اسنادی که در یک خوشه قرار دارند نسبت به پرسش رفتار (میزان ربط) مشابهی دارند.

در ارزیابی خوشه‌بندی سه دسته معیار قابل تمیز هستند: معیارهای بیرونی و معیارهای درونی و معیارهای نسبی.

همچنین بسیاری از الگوریتم‌های خوشه‌بندی، بر مبنای اصلاح پرسش توسط کاربر، با توجه به خوشه‌هایی که نمایش داده می‌شود و بازآرایی خوشه‌ها توسط نظام، به نتایج مطلوب می‌رسند. بر همین اساس، آنچه در رابط کاربر در معرض دید و انتخاب کاربر قرار می‌گیرد، نقش مهمی در بهبود نتایج بازیابی این نظام‌ها دارد. اگر رابط کاربر دارای ویژگی‌هایی چون سادگی و گویایی، انعطاف‌پذیری در انتخاب منابع و انواع قالب اطلاعات، کاربرپسند بودن (مانند تنظیمات و ترجیحات کاربر)، نمایش کلی و جزئی نتایج، نمایش گرافیکی و متنی جداگانه، و برخی ویژگی‌های تعاملی دیگر باشد، به‌نظر می‌رسد در بازیابی نتایج مرتبط با نیاز اطلاعاتی، نقش بیشتری داشته باشد.

این مطالعه، ضمن مرور مبانی نظری خوشه‌بندی، در پی شناسایی عناصر و مؤلفه‌های رابط کاربر موتورهای خوشه‌بندی بود. بر همین اساس، برخی عناصر پرکاربرد در ۱۰ موتور خوشه‌بندی مورد مطالعه، شناسایی شد. برخی از این عناصر، منحصر به موتورهای خوشه‌بندی است و برخی با موتورهای کاوش قبلی مشترک است، به همین دلیل، امروزه بهره‌گیری از موتورهای خوشه‌بندی

به عنوان مکمل جستجوی عبارتی و مروری در وب مورد تأیید است. توجه به عناصر مهم رابط کاربر چنین نظام‌های خوشه‌بندی که ویژگی‌های تعاملی و انعطاف‌پذیری را تقویت نماید، از نکات بنیادی موفقیت رابط کاربر چنین نظام‌هایی خواهد بود.

مطالعه مقاله حاضر و آشنایی با ویژگی‌ها و عناصر رابط کاربر موتورهای خوشه‌بندی می‌تواند سبب اندیشیدن پژوهشگران نرم افزاهای جستجوی وب به چگونگی طراحی رابط کاربر موتور خوشه‌بندی به زبان فارسی گردد.

۱۳. پیشنهادات

- می‌توان قابلیت تعامل رابط کاربر موتورهای خوشه‌بندی را با استفاده از قابلیت‌های گرافیکی و چندرسانه‌ای محیط وب، ارتقاء داد. اما، لازم است که میزان و چگونگی تعامل کاربر با عناصر در نظر گرفته شده در این گونه رابط کاربرها، مورد مطالعه قرار گیرد.
- تقویت ویژگی‌های نمایشی و گرافیکی نتایج بازیابی در نظام‌های خوشه‌بندی جنبه مهمی است که به راهنمایی کاربران به سمت جستجوی عمیق‌تر و منظم‌تر و ایجاد انگیزه تعامل بیشتر در آنها کمک خواهد کرد. لازم است در این زمینه، پس از مطالعات عمیق، به اصول و استانداردهای نمایش بصری در زبان فارسی، دست یابیم.
- خوشه‌بندی نتایج جستجو فقط مختص منابع وب نیست و می‌توان استفاده از آن را در نتایج جستجوی پایگاه‌های اطلاعاتی و پورتال‌ها و نرم‌افزارهای کتابخانه‌ای مورد بررسی قرار داد.
- با توجه به اینکه عناصر (۲۸ عنصر) رابط کاربر موتور خوشه‌بندی شناسایی شده در این مطالعه، مربوط به زبان انگلیسی است، جا دارد در مورد کاربردی بودن این عناصر در موتورهای خوشه‌بندی زبان فارسی هم مطالعات عمیقی صورت گیرد.
- مقایسه بین اهمیت عناصر رابط کاربرهای وب، به‌تنهایی می‌تواند موضوع مطالعات آینده قرار گیرد.
- علاوه بر رابط کاربر، جنبه‌های دیگر نظام‌های خوشه‌بندی از جمله الگوریتم‌های خوشه‌بندی، شیوه انتخاب برچسب و تعداد خوشه‌ها، استخراج توصیف برای نتایج و کاربرد اصطلاحنامه برای بسط یا محدود کردن دامنه جستجو، جای مطالعه بسیار دارد.

۱۴. منابع

ارسطوپور، شعله. ۱۳۸۸. دسته‌بندی نتایج جستجو بر مبنای ویژگی‌های مدارک و امکان‌سنجی استفاده از الگوریتم‌های خوشه‌بندی مختلف در سطح وب. *فصلنامه کتابداری و اطلاع‌رسانی* ۴۶ (۱۲): ۱۴۵-۱۷۴.

اسفندیاری مقدم، علیرضا، و زهره بهاری موفق. ۱۳۸۸. امکانات جستجو در فراموتورهای جستجو در وب: رویکردی مبتنی بر سیاهه وارسی. *فصلنامه علوم و فناوری اطلاعات* ۲ (۶۰): ۲۶۵-۲۸۸.

چنگیزی، امیر. ۱۳۸۷. اصول طراحی رابط کاربری. *فصلنامه ره‌آورد* ۲۴ (۴۱): ۲۴-۲۹.

حریری، نجلا. ۱۳۸۳. ربط در مدل سنتی و مدل‌های تعاملی بازیابی اطلاعات. *اطلاع‌شناسی* ۱ (۲): ۹۱-۱۱۳.

زره‌ساز، محمد و رحمت‌الله فتاحی. ۱۳۸۵. ملاحظات اساسی در طراحی رابط کاربر نظام‌های رایانه‌ای و پایگاه‌های اطلاعاتی. *فصلنامه کتاب* ۱۷ (۶۶): ۲۵۱-۲۶۸.

فتاحی، رحمت‌الله. ۱۳۸۳. تحلیل عوامل مؤثر بر نسبی بودن ربط در نظام‌های بازیابی اطلاعات. *اطلاع‌شناسی* ۱ (۲): ۷-۲۲.

نوکاریزی، محسن، رحمت‌الله فتاحی، و محمدرضا داورپناه. ۱۳۸۶. بررسی میزان و عوامل مؤثر بر قابلیت درک واژگان محیط رابط نرم‌افزارهای جامع کتابخانه‌ای فارسی. *فصلنامه مطالعات تربیتی و روان‌شناسی* ۳ (۱۷): ۵-۳۰.

Baeza-Yates, R., and B. Riberio-Neto. 1999. *Modern Information Retrieval*. New York, ACM Press.

Borlund, P, and P Ingwersen. 1998. Measures of relative relevance and ranked half-life: Performance indicators for interactive IRIn. *Annual ACM Conference on Research and Development in Information Retrieval SIGIR'98*: 324 – 331.

Carpineto, C., S.Osinski, G. Romano, , and D. Weiss, 2009. A survey of web clustering engines. *ACM Computing Surveys* 41 (3). http://doi.acm.org (accessed 06/08/2010).

Cunningham, S. J. et al. 1998. Applying connectionist models to information retrieval. http://www.cs.waikato.ac.nz/~ihw/ papers/97SJC-GH-JL-RB-IHW-Apply.pdf (accessed 20/10/2010).

Dietinger, T. et al. 1999. Targeted Information Retrieval. http://www.iicm.tugraz.at/ iicm_papers/targetedinformationretrieval_ collaboration/tir_is_99.html (accessed 05/07/2010).

Dietinger T. et al. 1998. Intelligent knowledge gathering and management as new ways of an improved learning process. Proc. WebNet 98, Orlando, Florida, 244-249. http://www.iicm.tugraz.at/iicm_papers/learningondemand/ICCE99_learning_on_ demand.html%3binternal&action=_help.action (accessed 15/09/2010).

Harper, M. F., S. Sen, and, D. Frankowskdi. 2007. Supporting Social Recommendations with Activity-Balanced Clustering. http://delivery.acm.org/10.1145/1300000/1297262/p165-harper.pdf?key1=1297262&key2=6518646521&coll =GUIDE&dl=GUIDE&CFID=58181297&CFTOKEN=94139946 (accessed 25/10/2010).

Hearst, Marti. A. 2009. *Search user interfaces*. London: Cambridge University Press. http://searchuserinterfaces.com (accessed 10/10/2010).

Jacso, Peter. 2007. Clustering search results. Part III: the synergy of metasearching and clustering. *Online Information Review* 31 (3). www.emeraldinsight.com1468-4527.htm (accessed 14/06/2010).

Jain, A. K. , M. N. Murty, and P. J. Flynn. 1999. Data Clustering: a review. *ACM Computing Surveys* 31 (3): 264 – 323.

Manning, C. D., P. Raghavan, and Hi Schutze. 2008. *An introduction to information retrieval*. Cambridge University Press. http://www.informationretrieval.org (accessed 19/05/2010).

Mizzaro, S. 1997. Relevance: the whole history. *Journal of the American Society for Information Science and Technology* 48 (9): 810-832.

Salton, G. 1991. Development in automatic text retrieval. *Science* 253 (5023): 974-980.

Saracevic, T. 1999. Information science. *Journal of the American Society for Information Science and Technology* 50 (12): 1051-1063.

Saracevic, T. 2007a. Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: Nature and Manifestations of Relevance. *Journal of the American Society for Information Science and Technology* 58 (13): 1915–1933,

Saracevic, T. 2007b. Relevance: a review of the literature and a framework for thinking on the notion in information science. Part III: Nature and Manifestations of Relevance. *Journal of the American Society for Information Science and Technology* 58 (13): 2126–2144.

Weiss, D. 2001. A clustering interface for web search results in Polish and English. Master thesis, Poznań University of Technology, Poland.

Xu, Rui, and D. Wunsch II. 2005. Survey of clustering algorithms. *IEEE Transaction on Neural Networks* 16 (3): 645-678.

Yoo, I., and X. Hu. 2006. A comprehensive comparison study of document clustering for a biomedical. Digital Library MEDLINE. JCDL '06. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. http://delivery.acm.org/ 10.1145/1150000/1141802/p220-yoo.pdf?key1=1141802&key2=6328646521&coll=GUIDE&dl=GUIDE&CFID=58181387&CFTOKEN=80180802 (accessed 17/12/2010).

Zamir, O., and O. Etzioni. 1999. Grouper: a dynamic clustering interface to Web search results. *Computer Networks* 31 (11-16): 1361–1374.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

Survey of User Interface Components in Clustering-based Information Retrieval Systems

Sayyed Mahdi Hosseini*
Ph.D Student in Library and Information Science,
Department of LIS, Ferdowsi University

Information Sciences & Technology

Iranian Research Institute
For Science and Technology
ISSN 1735-5206
eISSN 2008-5583
Indexed in LISA, SCOPUS & ISC
Vol.26 | No.3 | pp: 625-653
spring 2011

Abstract: Clustering search engines are used widely to help relevant information retrieval on the web, in parallel with keyword searching and browsing. User interface is important as the front line of systems to face to users and to use feedback to optimize user query. The aim of this article was to identify the important user interface components of clustering engines. The necessity of clustering usage in information retrieval systems, their theoretical background, algorithms, applications, and evaluation criteria have been discussed. According to web information retrieval literature, 10 clustering engines were identified and their user interface elements and application were described. Finally, the use rate of these elements in 10 clustering engines was illustrated.

Keywords: clustering, information retrieval systems, user interface, clustering systems, search engines

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

* Corresponding Author: smahdihosseini@gmail.com