شخصی سازی پروفایل علاقه کاربران وب بر اساس خوشه‌بندی داده‌های کلیک

زینب دهقانی[1]، کمال میرزایی[2]، محمدرضا ملاخلیلی میبدی[3]

[1] دانشجوی کارشناسی ارشد، دانشگاه آزاد اسلامی، واحد میبد، گروه مهندسی کامپیوتر، میبد، ایران
Zeinab.Dehghani@gmail.com

[2] عضو هیات علمی،دانشگاه آزاد اسلامی، واحد میبد، گروه مهندسی کامپیوتر، میبد، ایران
K.Mirzaie@Maybodiau.ac.ir

[3] عضو هیات علمی،دانشگاه آزاد اسلامی، واحد میبد، گروه مهندسی کامپیوتر، میبد، ایران
Mollakhalili@Maybodiau.ac.ir

چکیده

با توجه به گسترش روز افزون اطلاعات صفحات وب، بازیابی اطلاعات از بین انبوه داده‌ها اهمیت ویژه‌ای پیدا کرده است. رویکرد کلی موتورهای جستجوی موجود، با استفاده از محتوای صفحات و ساختار ارتباطی موجود بین آن‌ها، میزان ارتباط صفحات با پرسش کاربر را ارزیابی نموده و نتایج را برمی‌گرداند. فرآیند رتبه‌بندی بسترآگاه، با در نظر گرفتن زمینه و تحلیل بستری از پیشینه پرس‌وجوی کاربر، صورت گرفته و رتبه‌بندی با آگاهی و با در نظر گرفتن پیش زمینه قبلی کاربر، انجام می‌شود.

در این مقاله، به منظور استخراج مفاهیم مد نظر کاربر، روشها و پیش نیازها، تحلیل و بررسی شده و چارچوبی جهت دریافت و ثبت اطلاعات زمینه‌ای کاربر جهت تشکیل پروفایل مفهومی ارائه خواهد شد. با خوشه‌بندی پروفایل مفهومی ایجاد شده، مفاهیم پر تکرار استخراج شده و صحت نتایج با معیارهای ارزیابی بررسی شده است.

کلمات کلیدی

بازیابی اطلاعات ، بسترآگاهی، رتبه‌بندی بر اساس رفتار کاربر، گراف مفهومی، خوشه‌بندی.

۱- مقدمه

اینترنت و نقش فزآینده آن در توزیع و دسترسی به منابع اطلاعاتی، در دهه‌های اخیر، مشکلاتی را در یافتن اطلاعات مرتبط با نیازهای کاربران، به وجود آورده است. بازیابی اطلاعات به فن‌آوری و دانش پیچیده جستجو و استخراج اطلاعات، داده‌ها و فراداده‌ها در انواع گوناگون منابع اطلاعاتی مثل بانک اسناد، تصاویر و داده های وب گفته می‌شود. در شرایطی که بخواهیم نتایج یافته شده به نتایج مورد نظر کاربر نزدیک باشد، می‌توان رتبه‌بندی را بر اساس نظر کاربر انجام داد. بنابراین تشخیص واستخراج رفتار کاربران از اهمیت خاصی برخوردار است. فرآیند رتبه‌بندی بسترآگاه[1]، با در نظر گرفتن زمینه و تحلیل بستری از پیشینه پرس‌وجوی کاربر، صورت گرفته و رتبه‌بندی با آگاهی و با در نظر گرفتن پیش زمینه قبلی کاربر، انجام می‌شود. در این شیوه، مشخصات محتوای صفحات را با رتبه‌بندی‌ها و قضاوت‌های کاربر ادغام نموده نتایج قابل قبول‌تری را ارائه می‌دهد.

جهت روشن شدن مطلب فرض کنید کاربری پرس‌وجویی با عنوان "قلعه حیوانات" را وارد نماید، موتور جستجو با تطابق این مورد با الگوهای موجود می‌تواند خروجی‌های زیر را تولید کند. مشخصات اکوسیستم‌های جانوری و یا مشخصات کتابی با این عنوان. چنانچه موتور جستجو از رویکرد محتوایی استفاده کند، بسته به کاربرانی که مشابهت رفتاری بیشتری با این درخواست دارند، هریک از دو پاسخ ممکن است به کاربر پیشنهاد شود. اما چنانچه موتور جستجو از رویکرد بسترآگاه برای بازیابی اطلاعات استفاده کند،

Archive of SID

www.SID.ir

نتیجه به زمینه و یا به عبارتی به پرس‌وجوی قبلی کاربر وابسته است. در این مقاله، به منظور استخراج مفاهیم مد نظر کاربر، روشها و پیش نیازها، تحلیل و بررسی شده و چارچوبی جهت دریافت و ثبت اطلاعات زمینه‌ای کاربر و تطابق آن با الگوهای محتوایی ارائه خواهد شد. در بخش دوم مقاله به مرور کارهای انجام شده در این زمینه می‌پردازد. در بخش سوم به منظور تبیین و تشریح ابعاد مساله، چارچوب پیشنهادی و اجزای آن بیان شده است و در ادامه چگونگی استخراج مفاهیم مرتبط در ایجاد گراف مفاهیم بررسی شده است. در بخش چهارم و پنجم ضمن تشریح روش خوشه‌بندی، نتایج تجربی خوشه‌بندی گراف مفهومی ارائه شده است. در پایان ضمن مرور مطالب بیان شده به جمع‌بندی و نتیجه‌گیری پرداخته شده است.

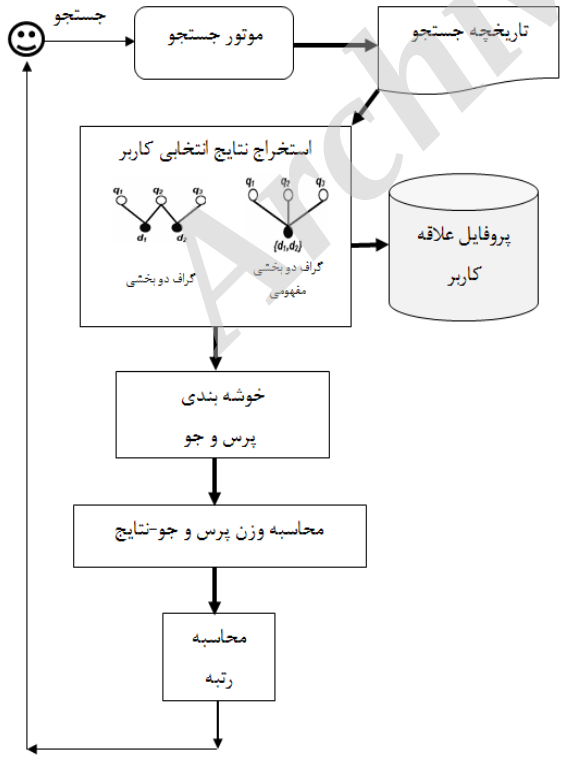
۲- مروری بر کارهای انجام شده

رتبه‌بندی یکی ازمهمترین قسمتهای موتورجستجو است. رتبه‌بندی یکی از مولفه‌های تکمیل کننده در بازیابی اطلاعات است[۱]. پیدا کردن صفحات مرتبط به هم یکی از کارکردهای رتبه‌بندی صفحات وب می‌باشد. در واقع رتبه‌بندی صفحات وب می‌تواند ساختار ارتباطی صفحات وب را نمایان سازد. رتبه‌بندی به سه روش اصلی تقسیم‌بندی شده است[۲]. روشهای مبتنی بر محتوا و روشهای مبتنی براتصال و روشهای رتبه‌بندی بر مبنای رفتار کاربر. بیشتر الگوریتم‌های رتبه‌بندی روش‌های مبتنی بر محتوا و مبتنی بر اتصال به صورت برون‌خط می‌باشند. بنابراین در بسیاری از موارد نتایج بازگردانده شده از موتورهای جستجو با انتظارات کاربر منطبق نمی‌باشد. در حالت بهتر می‌توان رتبه‌بندی را بر اساس نظر کاربر انجام داد. بدیهی است که در این روش علاوه بر پرس‌وجوها، رفتارهای کاربران نیز مورد بررسی قرار گرفته و بخشی از ورودی الگوریتم‌های رتبه‌بندی به حساب می‌آیند. بنابراین تشخیص و استخراج رفتار کاربران از اهمیت خاصی برخوردار می‌شود. در سال‌های اخیر، توجه به رفتار کاربر به عنوان یکی از مهمترین معیارهای رتبه‌بندی، مورد توجه واقع شده است. نزدیک بودن نتایج رتبه‌بندی به نیازها و اولویت‌های کاربران اصلی‌ترین هدفی است که در [۳] به آن پرداخته شده است. در این مرجع، سه روش جهت کشف و استخراج رفتار کاربر بیان نموده است. ۱) بازخورد صریح[۲] و رویکرد ضمنی. ۲) ایجاد پروفایل کاربر بر اساس تاریخچه جستجوها ۳) ایجاد پروفایل کاربر به صورت فیلترکردن جمعی(مشارکتی)[۳]. در هر سه روش پروفایلی برای کاربر ساخته می‌شود. هر بار که کاربر پرس‌وجویی را درخواست می‌کند نتایجی توسط موتور جستجو به کاربر برگردانده می‌شود. کاربر از بین نتایج تعدادی را انتخاب نموده و صفحات مربوط را بازدید می‌کند. نتیجه بازدیدها و انتخاب‌های کاربر در پروفایل ثبت شده و با بروز رسانی پروفایل کاربر تاریخچه رفتار کاربر را نگه می‌دارد. نکته جالب توجه این است که برای ساخته شدن این پروفایل، هیچ تلاشی متوجه کاربر نبوده است. جهت ایجاد پروفایل کاربر بر اساس تاریخچه جستجوها، اولویت‌ها و علائق کاربر به دو گروه اولویت‌های دائمی و بلند مدت و اولویت‌های موقتی یا کوتاه مدت ، دسته‌بندی می‌شوند. ساختن پروفایل کاربر بر اساس تاریخچه جستجوها، با ثبت و پی‌گیری نشست‌های کاربر صورت می‌گیرد. می‌توان برداری از پروفایل‌های کاربر در مراجعات مختلف ساخت. برای هر کاربر پروفایلی با عنوان P در نظر گرفته شده است. پروفایل دائمی کاربر به صورت P_{pre} و پروفایل موقتی کاربر را با P_{today} نمایش داده می‌شود. P_{pre} در واقع نمایانگر پروفایل‌های کاربر در n روز گذشته است. تاریخچه مراجعات و وضعیت پروفایل کاربر را در امروز و n روز گذشته قابل نگهداری است. در این روش مفهوم اندازه پنجره برای نشان دادن P_{pre} به کار رفته است. همچنین $S_j\ (j=0,1,2,...,n)$ تعداد صفحات وب است که کاربر در j امین روز مشاهده کرده است. بنابراین j=۰ نمایش دهنده امروز است. نشست فعلی کاربر در امروز با cur نشان داده شده است. n_{bh} نشان دهنده جستجوهای مختلف کاربران قبل از موقعیت cur در امروز بوده است. بنابراین رابطه بین n_{bh} و cur به صورت زیر است:cur= n_{bh}+۱. بنابراین پروفایل P_{today} هر روز به همین صورت ساخته می‌شود.

Storey در[۴] با ارائه متدولوژی CONQUER[۴] پردازش پرس‌وجو آگاه از متن، با به کارگیری دو منبع دانش تکمیلی شامل واژگان و هستی شناسی، محتوای معنایی نمایش داده شده حاصل از پرس‌وجوی کاربر را افزایش می‌دهد. این متدولوژی از پرس‌وجو به عنوان یک دانه ، استفاده نموده و با ساخت یک شبکه معنایی، با به کارگیری دو منبع دانش و تطابق آنها، شبکه ساخته شده را تصحیح می‌نماید. پرس‌وجوی نشان داده شده توسط شبکه معنایی تصحیح شده، می‌تواند توسط موتورهای جستجو اجرا شود. تجربه‌های آزمایشگاهی نشان داده است که پرس‌وجوی تصحیح شده حاصل از متدولوژی، نتایج بهتری را در بر داشته است. رویکرد کلی این متدولوژی بر مبنای پردازش پرس‌وجو است.

۳- چارچوب پیشنهادی بر اساس پروفایل کاربر

در این بخش روش پیشنهادی جهت بهبود مدل رفتاری کاربر ارائه شده است. نخست ضمن تعریف دقیق بیان مساله، معماری سیستم ارائه شده و مولفه‌های آن و نحوه‌ی عملکرد آن‌ها توضیح داده می‌شود. شکل(۱) معماری چارچوب پیشنهادی را نشان می‌دهد.



شکل(۱). چارچوب پیشنهادی مبتنی بر پروفایل کاربر

Archive of SID

www.SID.ir

در چارچوب پیشنهادی ، ارائه‌ی روشی برای بهبود مدل رفتاری کاربر با توجه به واکنش‌ها و رفتارهای قبلی است که در صورت تعمیم و گسترش آن، می‌تواند ویژگی های زیر را داشته باشد:

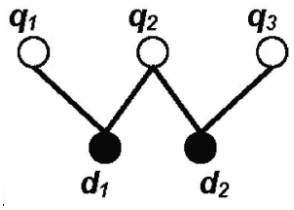
- مدل بصورت ضمنی یعنی بدون دخالت مستقیم کاربر ساخته شود.
- مدل بصورت فردی باشد، یعنی به ازای هر کاربر یک مدل خاص وجود داشته باشد.
- مدل براساس رفتار گردشی کاربر در بازه‌ی خاصی از زمان مثلا سه ماه گردش او در وب‌سایت ساخته شود.
- فرایند ساخت مدل تا حد امکان بصورت خودکار باشد.

فرض کنید کاربر U در بازه‌ی زمانی T از صفحات وب‌سایت دیدن کرده است و نشست‌های $\{s_1,s_2,\ldots,s_m\}$ را داشته است. هدف، ساختن یک بردار از سابقه جستجوهای آن کاربر است، برداری که پروفایل کاربر را تشکیل دهد. همچنین چنانچه کاربر پرس‌وجوی Q را در وب انجام دهد، پرس‌وجوی او نیز در پروفایل وی جز سابقه پروفایل وی محسوب گردد. مجموعه انتخاب‌های کاربر نیز به پروفایل کاربر اضافه شود. به عبارتی برای هر کاربر بردار $(V=<(q_1,c_1,w_1),(q_1,c_2,w_2),\ldots,(q_1,c_n,w_n)>)$ ، بطوری که q نشان‌دهنده پرس‌وجوی کاربر و c نشان‌دهنده نتایج انتخابی باشند وw وزن‌های آن‌ها نیز تا حد امکان نشان‌دهنده‌ی اهمیت آن‌ها در رفتار کاربر در آن بازه‌ی زمانی باشد.

۳-۱- استخراج نتایج انتخابی کاربر

۳-۱-۱- ساخت گراف دو بخشی مفهومی

با دنبال کردن کلیک‌های کاربر، گراف دو بخشی ساخته می‌شود(شکل(۲)) [۵]. مجموعه پرس‌وجوهای کاربر با q و مجموعه اسنادی که به عنوان نتایج پرس‌وجوی کاربر برگردانده شده است با d نشان داده می‌شود.



شکل(۲). گراف دو بخشی سند-پرس و جو

گراف دو بخشی سند- پرس‌وجو، در واقع نمای دیگری از دنباله کلیک‌های کاربر است. زمانی که کاربر بر روی سندی کلیک کند یالی بین پرس‌وجو و سند ایجاد می‌شود. بدیهی است که کاربر می‌تواند برای یک پرس‌وجو، چندین سند را مشاهده نماید. با داشتن گراف دو بخشی امکان دسته‌بندی مفاهیم فراهم می‌شود. الگوریتم ۱، ساخت گراف دو بخشی را نشان می‌دهد.

الگوریتم۱. ساخت گراف دو بخشی

ورودی: دنباله کلیک از گذر داده(CT)، مجموعه مفاهیم(E)

خروجی : گراف دو بخشی مفهومی(G)

۱. به دست آوردن مجموعه پرس‌وجوهای مشابه $Q=\{q_1,q_2,q_3,\ldots\}$ از CT.

۲. به دست آوردن مجموعه مفاهیم مشابه $C=\{c_1,c_2,c_3,\ldots\}$ از E.

۳. ساختن گره های گراف G از اشتراک دو مجموعه Q و C $(Q\cap C)$ به طوری که Q و C در طرفین گراف G باشند.

۴. به ازای هر $q_i \in Q$ چنانچه کاربر بر روی مفهوم C_j کلیک کرده باشد، یال $e=(q_i, c_j)$ به گراف G اضافه شود.

۳-۱-۲- استخراج مفاهیم

استخراج مفاهیم و شباهت‌های اسناد و پرس‌وجوهای مرتبط می‌تواند پارامتر موثری در فرایند خوشه‌بندی باشد. روش‌های مختلفی جهت استخراج مفاهیم معرفی شده است: استخراج مفاهیم با استفاده از بخش‌بندی وب حاصل از نتایج برگردانده شده موتور جستجو،کاوش در مفاهیم مشابه و ساخت پروفایل مفهومی کاربر.

دو روش اول بر اساس مشابهت ترمها و پرس‌وجوها عمل می‌کند. مبنای روش بخش‌بندی وب بر اساس نتایج حاصل از پرس‌وجوها است. با مرور یک صفحه وب توسط کاربر، لیستی از نتایج را در پیش رو دارد.

روش سوم بر مبنای گراف ارتباط مفهومی پایه‌گذاری شده است. گراف ارتباط مفهومی از روی داده‌های به دست آمده از کلیک کاربران ساخته می‌شود. این گراف، مفاهیم قابل برداشت از پرس‌وجوی کاربران را نشان می‌دهد. برای روشن شدن مطلب، چگونگی ساخته شدن گراف مفهومی را با یک مثال بیان می‌کنیم. فرض کنید کاربر پرس‌وجوی "قلعه حیوانات" را درخواست نماید، با دریافت این پرس‌وجوی، فضای مفهومی از روی بخش‌های وب مرتبط با مفاهیمی چون "جرج اورل"، "۱۹۸۴" و "فروشگاه کتاب" ساخته می‌شود. اگر کاربر به موضوعات مربوط به کتاب علاقه داشته باشد و مفهوم "فروشگاه کتاب" را کلیک کند، داده‌های مربوط به کلیک کاربران به تدریج به این مفهوم، نزدیک شده و به وزن آن گره اضافه نموده و جز همسایه‌های آن گره محسوب می‌گردد. اما وزن مفاهیم غیرمرتبط مانند "اکوسیستم‌های جانوری" و "گونه های جانوری" و ... صفر باقی می‌ماند. بنابراین با هر بار کلیک مفاهیم مرتبط، یکی به وزن گره متناظر آن اضافه می‌شود(رابطه (۱) و(۲)).

$$click(s_j) => \forall t_i \in s_j, W_{t_i} = W_{t_i} + 1 \quad (۱)$$

$$click(s_j) => \forall t_i \in s_j, W_{t_i} = W_{t_i} + sim_R(t_i,t_j) if sim_R(t_i,t_j) > 0 \quad (۲)$$

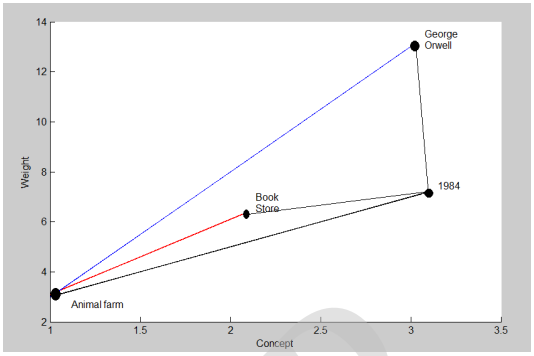
در رابطه (۱)، s_j نماینده بخشی از وب است که توسط کاربر کلیک می‌شود و با $click(s_j)$ نمایش داده می‌شود. t_i نشان دهنده هر یک از مفاهیم و t_j همسایه t_i است. وزن مفاهیم با W_{ti} نمایش داده می‌شود.

زمانی که کاربر بر روی مفهوم s_j کلیک می‌کند، وزن مفهوم t_i در s_j ظاهر می‌شود و جهت منعکس نمودن علائق مفهوم موجود در s_j، یک واحد به وزن آن اضافه می‌شود. اگر مفهومی به مفهوم کلیک شده نزدیک باشد به

Archive of SID

www.SID.ir

وزن آن یک واحد اضافه شده در غیر این صورت تنها کسری از یک (که حتی ممکن است به صفر میل کند) اضافه می‌شود. برای دریافت علائق کاربر، پروفایل مفهومی علائق کاربر برای پرس‌وجوی ورودی ساخته می‌شود. شکل (۳) نمونه‌ای از گراف ارتباط مفهومی بین مفاهیم مشابه را نشان می‌دهد.



شکل (۳). گراف ارتباط مفهومی

٤- خوشه‌بندی گراف مفهومی

خوشه‌بندی مشاهده صفحه‌ها و پرس‌وجوها می‌تواند به عنوان یکی از رویکردهای کاوش داده‌های وب، مورد تحلیل واقع شود. با داشتن یک مجموعه از پرس‌وجوها، تکنیک‌های متنوعی جهت کشف دانش بدون ناظر می‌توانند برای استخراج شباهت‌ها و مفاهیم بکار روند. الگوریتم های استاندارد خوشه‌بندی مانند k-means با در نظر گرفتن یک معیار شباهت یا فاصله گراف پرس‌وجو–اسناد را به گروه‌هایی تقسیم می‌کنند. خوشه های بدست آمده از این روش می‌توانند گروه‌هایی از کاربران را بر مبنای رفتار گردشی آن‌ها نمایش دهند. روش خوشه‌بندی k-means، نقاطی به عنوان مراکز خوشه‌ها به عنوان میانگین نقاط هر خوشه در نظر گرفته و سپس هر نمونه داده را به یک خوشه با کمترین فاصله تا مرکز آن خوشه نسبت می‌دهد. در برآورد خوشه‌ها تابع رابطه (۳) به عنوان تابع هدف در نظر گرفته شده است.

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (۳)$$

فاصلهٔ بین نقاط با عبارت داخل $\| \ \|$ سنجیده می‌شود و c_j مرکز خوشهٔ j ام است. بدیهی است در صورت کم بودن میزان فاصله، نقطه x_i درخوشه با مرکز c_j قرار می‌گیرد.

٥- آزمایش‌های تجربی

٥-١- مجموعه داده

جهت بررسی و ارزیابی استخراج مفاهیم پرس‌وجوهای مشابه، بخشی از داده‌های کلیک داده گروه تحقیقاتیAOL که به صورت منبع باز در سال ۲۰۰۶ منتشر شده، استفاده شده است. این داده‌ها در ۱۰ دسته فایل متنی به صورت جداگانه قرار دارد که بر روی سرور *Apache/۲.۲.۱۴* در دسترس است [۶]. در این پژوهش از فایل متنی گروه یک از گروههای ده‌گانه استفاده شده است. جدول ۱ اطلاعات آماری مجموعه‌داده منتشر شده توسط AOL را نشان می‌دهد.

جدول ۰۱ اطلاعات آماری مجموعه‌داده AOL

تعداد رکوردها	۳۶,۳۸۹,۵۶۷
تعداد نمونه از هر پرس‌وجو	۲۱,۰۱۱,۳۴۰
تعداد کل صفحات انتخاب شده	۱۹,۴۴۲,۶۲۹
تعداد پرس‌وجوی نرمال شده	۱۰,۱۵۴,۷۴۲
تعداد شناسه کاربر	۶۵۷,۴۲۶

هر رکورد در فایل محتوی پنج فیلد بوده که شامل AnonID, Query, QueryTime, ItemRank, ClickURL است. با داشتن پرس‌وجو و صفحه انتخابی کاربر، برای هر کاربرگراف پرس‌وجو–سند، ساخته می‌شود. دو فیلد Query و ClickURL به ترتیب پرس‌وجوی کاربر و صفحات انتخاب شده به ازای کلیک کاربر روی نتایج برگردانده شده پرس‌وجو را نشان می‌دهد. رکوردها بر اساس فیلد AnonID که شماره شناسایی کاربر است، مرتب شده‌اند. وجود فیلد AnonID، امکان شناسایی و تمایز کاربران را فراهم کرده است.

پس از نرمال‌سازی بر اساس دو فیلد پرس‌وجو و صفحات انتخابی کلیک شده، از بین کاربران، ۶۵ کاربر به تصادف انتخاب شده و برای هر یک پروفایل جستجوی جداگانه ایجاد شده است. بدیهی است که بسته به مدت زمان و تعداد دفعات جستجوی هر کاربر، تعداد رکوردهای پروفایل‌های کاربران انتخابی با یکدیگر متفاوت بوده اند. حداقل تعداد رکوردهای پروفایل که نمایانگر تعداد درخواست‌های پرس‌وجوی کاربر به موتور جستجو است، ۲۹ و حداکثر تعداد رکوردها ۴۹۴ است. جدول ۲ تعداد پرس‌وجوی ۶ کاربر از ۶۵ کاربر انتخابی را نشان می‌دهد.

جدول ۲. تعداد پرس‌وجوهای کاربران انتخابی

شماره شناسایی کاربر	تعداد پرس‌وجو(رکورد)
۲۳۳۴	۱۶۱
۳۳۰۲	۲۹۸
۳۷۴۵	۱۶۴
۴۷۸۱	۴۱۳
۱۱۱۷۳۵	۲۷۰
۱۱۲۲۴۰	۴۶۴
کاربر با کمترین پرس‌وجو	۲۹
کاربر با بیشترین پرس‌وجو	۴۹۴

٥-٢- نتایج شبیه‌سازی

با استفاده از ابزارهای تشخیص و استخراج کلمات کلیدی، واژه‌های کلیدی اسناد و پرس‌وجوهای هر کاربر به طور جداگانه، خارج شده و پروفایل مفهومی کاربران از روی پرس‌وجوها، اسناد و وزن پرس‌وجو–سند ساخته شده است. به ازای هر پرس‌وجو و سند انتخابی، وزن تعداد دفعات انتخاب اسناد، محاسبه و انتساب داده شده است که گراف مفهومی هر کاربر را تشکیل می‌دهد. جدول ۳ اسناد انتخابی کاربر با شماره شناسایی ۲۱۷۸ به ازای چند پرس‌وجوی یکسان و تعداد دفعات انتخاب اسناد را نشان می‌دهد.

Archive of SID

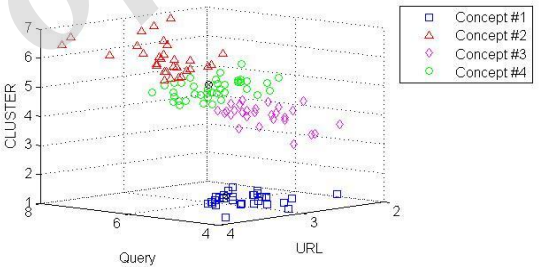
www.SID.ir

جدول ۳. گراف مفهومی کاربر ۲۱۷۸

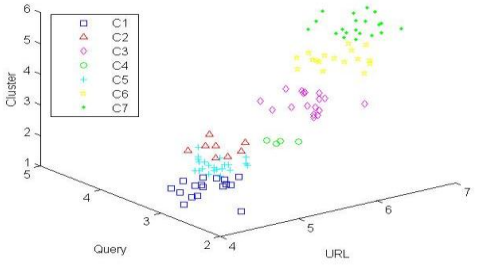
دفعات بازدید	سند انتخاب شده	پرس‌وجو
۲	'www.longbeachmuffler.com'	'honda accord fuel additives check engine light'
۱	'http://www.edmunds.com'	'honda accord fuel additives check engine light'
۲	'www.washingtonpost.com'	'honda accord fuel additives check engine light'
۲۳	'http://experts.about.com'	'honda accord fuel additives check engine light'
۵	'http://www.alldata.com'	'honda accord fuel additives check engine light'
۲	'http://townhall-talk.edmunds.com'	'honda accord fuel additives check engine light'
۲۵	'www.homeportfolio.com'	'pergola'
۲۸	'http://www.plansnow.com'	'pergola'

با توجه به گراف جدول ۳ می‌توان تعداد تراکم بازدید یک صفحه را تشخیص داد.

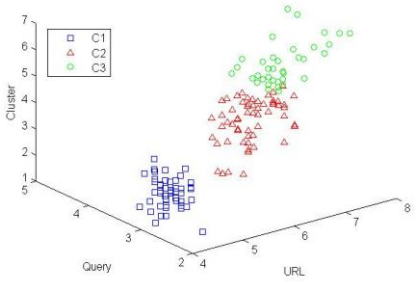
هدف این مقاله، شبیه‌سازی استخراج صفحاتی است که به ازای یک پرس‌وجوی کاربر بیشترین مراجعه به آن صفحه را داشته است. به عبارتی از روی میزان تراکم بازدید یک صفحه، می‌توان به موضوعات مورد علاقه کاربر بر اساس مشاهدات قبلی وی، پی برد. اعمال روش‌های خوشه‌بندی بر روی گراف مفهومی حاصل از جدول ۳، پرس‌وجوها و صفحات بازدید شده را برای پروفایل هر کاربر دسته‌بندی می‌کند. شکل (۴) گراف مفهومی حاصل از سه پروفایل سه کاربر را نشان می‌دهد.



الف. خوشه‌بندی مفهومی کاربر با ۲۹۸ رکورد



ب. خوشه‌بندی مفهومی کاربر با ٤١٣ رکورد



ج. خوشه‌بندی مفهومی کاربر با ١٦٤ رکورد

شکل (٤). خوشه‌بندی مفهومی پروفایل ۳ کاربر

مطابق با شکل (۴) الف از پروفایل کاربر۴ مفهوم استخراج شده است. بدیهی است که با افزایش تعداد رکوردهای سوابق پروفایل کاربران، تعداد مفاهیم استخراج شده افزایش می‌یابد. به طوری که در شکل (۴) ب که کاربر جستجوهای بیشتری انجام داده است، تعداد ۷ مفهوم استخراج شده است. در شکل (۴) ج پروفایل کاربر به ۳ خوشه تقسیم شده است.

۵-۳- ارزیابی نتایج

تکنیک اعتبارسنجیSilhouette[۵] یکی از روشهای اعتبارسنجی خوشه‌ای است که در ارزیابی کلاس‌های خوشه‌بندی مورد استفاده قرار می‌گیرد. مقدار Silhouette از رابطه (۴) محاسبه می‌شود.

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (۴)$$

a(i) میانگین فاصله بین مشاهده i با سایر مشاهدات در یک خوشه مشابه و b(i) میانگین فاصله مشاهده i به تمام مشاهدات در خوشه‌های دیگر می باشد. بنابراین برای پارامتر ارزیابی کننده s(i) داریم:

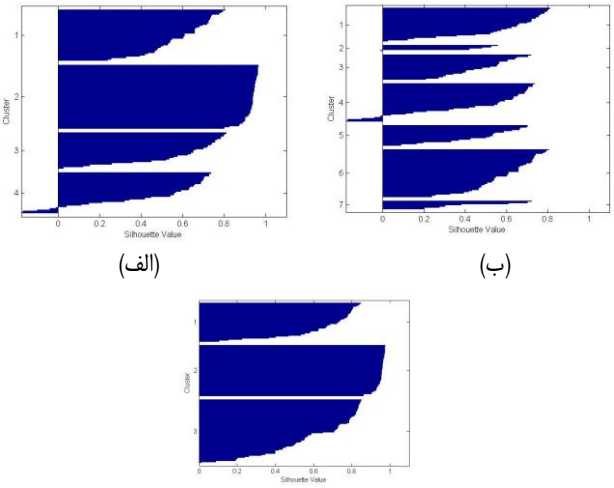
$$S(i) = \begin{cases} 1 - a(i)/b(i) & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1 & \text{if } a(i) > b(i) \end{cases} \qquad (۵)$$

در خوشه‌بندی بهینه تقسیم‌بندی مجموعه داده‌ها به گونه‌ای است که داده‌های موجود در هر گروه بیشترین شباهت را با یکدیگر و بیشترین تفاوت را با داده‌های موجود در گروه‌های دیگر دارند. مقدار s(i) در رابطه (۴) بین ۱+ و ۱- قرار دارد. نزدیک بودن مقدار s(i) به ۱+ نمایانگر خوشه‌بندی مناسب نمونه‌ها و نزدیک بودن مقدار s(i) به ۱- نمایانگر خوشه‌بندی نامناسب نمونه‌ها است [۷].

میزان پارامتر Silhouette برای سه خوشه‌بندی شکل (۴) محاسبه شده است(شکل(۵)). بهترین پاسخ برای این پارامتر وقتی است که به ۱+ نزدیک باشد.

Archive of SID

www.SID.ir



(الف) (ب)

(ج)

شکل(۵). پارامتر Silhouette برای خوشه‌بندی سه کاربر انتخابی

همانطور که در شکل(۵) دیده می‌شود، مقدار Silhouette برای هر سه کاربر به ۱+ نزدیک است که نشان دهنده خوشه‌بندی مناسب مفاهیم است.

٦- بحث و نتیجه‌گیری

با داشتن دنباله کلیک از گذر داده، گراف دو بخشی ساخته می‌شود. همچنین با استخراج مفاهیم از یک دنباله کلیک از گذر داده و یا گراف دو بخشی پرس‌وجو - اسناد گراف کلی‌تری ایجاد می‌شود. این گراف دو بخشی مبتنی بر مفاهیم است. الگوریتم ۱ نحوه ساخت یک گراف دو بخشی مبتنی بر مفاهیم را از دنباله کلیک نشان می‌دهد. این الگوریتم با درنظر گرفتن دنباله کلیک‌های کاربر و با داشتن مفاهیم استخراج شده، گراف دو بخشی مفهومی را ایجاد می‌کند. گراف دو بخشی مفهومی یک گام از یک گراف دو بخشی پرس‌وجو - سند جلوتر بوده و در ایجاد پروفایل مفهومی نقش مهمی را ایفا می‌کند. با به کار گرفتن گراف دو بخشی و دنبال کردن مشابهت صفحات انتخابی کاربر به ازای پرس‌وجوها، می‌توان ارتباط مفهومی پرس‌وجوها را کشف نمود. در واقع گراف مفهومی نشان دهنده پرس‌وجوهایی است که شباهت بیشتری دارند. گراف مفهومی زمینه‌ای برای ساخت پروفایل مفهومی کاربر را فراهم می‌کند. خروجی گراف مفهومی در خوشه‌بندی پرس‌وجوها استفاده می‌شود. با اعمال روش‌های خوشه‌بندی می‌توان شباهت بین خوشه‌ها را استخراج نمود. در طول فرایند خوشه‌بندی، پرس‌وجوهای مشابه با یکدیگر ترکیب می‌شوند. در این مقاله با استخراج شباهت بین اسناد و پرس‌وجوهای کاربران، میزان همبستگی آنها تعیین شده است. جهت بررسی و ارزیابی استخراج مفاهیم پرس‌وجوهای مشابه، بخشی از داده‌های کلیک داده گروه تحقیقاتیAOL، استفاده شده است. از روی داده‌های کلیک برای هر کاربر پروفایل مفهومی ارائه شده است. با خوشه‌بندی پروفایل مفهومی ایجاد شده، مفاهیم پر تکرار استخراج شده و صحت نتایج با معیارهای ارزیابی بررسی شده است.

مراجع

[۱] Brodin, A., Robers,G.O,Rosenthal,J.S,and Tsaparas,P., Link Analysis Ranking:Algorithms,Theory,and Experiments. ACM Transactions on Internet Technology, February ۲۰۰۵. ۵(۱): p. ۲۳۱-۲۹۷.

[۲] Bidoki, A.M.Z., Effective Web Ranking & Crawling, in School of Electrical and Computer Engineering. May ۲۰۰۹, University of Tehran: Tehran.

[۳] Kazunari Sugiyama, K.H., Masatoshi Yoshikawa. Adaptive Web Search Based on User Profile Constructed without Any Effort from Users. in Proceedings of the ۱۳th international conference on World Wide Web ۲۰۰۴. New York, NY, USA

[٤] Veda C. Storey, A.B.-J., Vijayan Sugumaran,Sandeep Purao, A Methodology for Context-Aware Query Processing on the World Wide Web. InformationSystems Research, March ۲۰۰۸. ۱۹(۱): p. ۳–۲۵.

[٥] KWT Leung, W.N., DL Lee, Personalized Concept-Based Clustering of Search Engine Queries. Knowledge and Data Engineering, IEEE Transactions, ۲۰۰۸. ۲۰(۱۱): p. ۱۵۰۵ - ۱.۵۱۸

[٦] G. Pass, A.C., C. Torgeson, A Picture of Search, in The First International Conference on Scalable Information Systems. June ۲۰۰۶: Hong Kong.

[۷] Tan P-N., “Steinbach M. and Kumar V. Introduction to Data Mining”, Pearson Addison Wesley, pp. ۷۶۹, ۲۰۰۶.

زیرنویس‌ها

۱ Context-Aware
۲ Relevance Feedback
۳ Collaborative filtering
٤ CONtext-aware QUERy processing
٥ Silhouette Validation Technique

Archive of SID

www.SID.ir