اولین همایش منطقه ای رویکردهای نوین در مهندسی کامپیوتر و فناوری اطلاعات
۵ خرداد ۱۳۹۰ - رودسر

مروری بر شخصی سازی نتایج موتور جستجو با روشهای هوشمند

دکتر مهدی یعقوبی[1]،ملیحه محمدزاده[2]

[1]دانشکده فنی مهندسی،دانشگاه آزاد اسلامی واحدمشهد
yaghoubi@mshdiaum.ir
[2] دانشکده فنی مهندسی،دانشگاه آزاد اسلامی واحد مشهد
malihemohamadzade@yahoo.com

چکیده

با افزایش روز افزون اطلاعات در وب ، یافتن مطالب مورد نیاز امر دشوار و وقت گیری شده است. شخصی سازی نتایج موتور جستجو هر تلاشی است که باعث شود نتایج جستجو متنا سب با علایق و دانش ونیاز کاربر نمایش داده شود.
شده است که کارهای انجام شده در حیطه شخصی سازی نتایج موتور جستجو از دو منظر نحوه انجام کار و مدل جمع آوری اطلاعات بررسی شده اند.در این مقالات از دو دیدگاه فازی و قطعی با مساله برخورد شده است و روشهای موجود بررسی شده است.

کارهای انجام شده در این محدوده به ویژه در سالهای اخیر به موفقیتهای بسیاری رسیده اند و توانسته اند نتایج موتورهای جستجو را تا سطح زیادی بهبود بدهند.هر چند هنوز هیچ سیستم تمام اتوماتیک و تمام فازی که در آن از مفهوم کاوی صفحات وب استفاده شود،ایجاد نشده است.

کلمات کلیدی

شبکه های مفهومی فازی ، شخصی سازی ،مدل سازی کاربر ، موتور جستجو

۱ مقدمه

درشخصی سازی نتایج موتور جستجو سعی می شود تا نتایجی که موتورهای جستجو نمایش می دهند با علایق و دانش و نیازهای کاربر ومنظور کاربر از جستجو متناسب باشد.یعنی می خواهیم ترتیب نتایجی که در یک موتور جستجو بر اساس پرس و جوی کاربر نمایش داده می شود به ترتیب علاقه کاربر باشد.

در این مقاله کارهایی که در حیطه شخصی سازی نتایج موتورهای جستجو انجام شده بررسی شده است روشی که در همه این کارها مشترکا انجام شده است این است که ابتدا کاربر شناخته می شود و نیازها و حیطه کاری او تشخیص داده می شود و سپس سیستم بر اساس شناختی که از کاربر دارد نتایج جستجو را مرتب میکند.بر این اساس ما کارهای انجام شده را از دو منظر نحوه مدل کردن کاربر(شناختن کاربر) و نحوه اجرای سیستم شخصی سازی بررسی کرده ایم.در قسمت دوم کارهای انجام شده را از منظر نحوه مدل کردن کاربر بررسی کرده ایم.در قسمت سوم نیز کارهای انجام شده را از منظر نحوه اجرای سیستم بررسی کرده ایم.در قسمت سوم مزایا و معایب کارهای انجام شده را بررسی می کنیم.در قسمت چهارم نتیجه گیری و در قسمت پنجم مراجع آورده شده است.

۲ کارهای انجام شده از منظر نحوه مدل سازی کاربر

در زمینه نحوه جمع آوری اطلاعات کاربر دو رویکرد وجود دارد.

2-1-مدل سازی صریح کاربر

در رویکرد اول اطلاعات شخصی کاربر به صورت صریح از او پرسیده می شود و بر اساس پاسخهایی که کاربر به سیستم اعلام می کند پروفایل شخصی او ساخته میشود.مثل کارهای انجام شده در[29],[2],[3],[4],[5],[6],[12],[14] و همچنین personal google.

در [2] از یک کلاسترینگ سلسله مراتبي استفاده مي كند كه snaket ناميده مي شود كه نتايج حاصل از 16 موتور جستجوي كالا در پوشه هاي بر چسب زده سلسله مراتبي مرتب مي شوند. حالت سلسله مراتبي ديد كاملي از نتايج مرتب شده موتورهاي جستجو ارائه مي دهد.كه كاربران با اين حالت سلسه مراتبي به نيازهاي جستجوي خودشان هدايت ميشوند.ابتدا كاربر يک پرس وجو را به SNAKET مي فرستد و SNAKET نيز پس از خوشه بندی سلسله مراتبی نتايج جستجو و برچسب گذاري خوشه ها با جملات با طول متغير، يک سلسله مراتب برچسب گذاري شده را به كاربر ارائه ميدهد. كاربر نيز گروههايي را كه برچسب آنها بيشترين تناسب با اطلاعاتی مورد نيازش را دارد انتخاب ميكند .سپس SNAKET با فيلتركردن نتايج جستجوي متعلق به سايرخوشه ها، نتايج جستجوي شخصي سازي شده را به كاربر ارائه مي دهد.

همانطور که دیده می شود این روش به دخالت صریح کاربر نیاز دارد که روش مطلوبی نیست.

در [4] از یک پایگاه دانش استفاده شده که از رفتار کاربران به دست آمده است سیستم رفتار کاربران را مانیتور می کند سپس کاربر خودش حیطه علایقش را معین می کند و بر این اساس مدل علایق کاربر به دست می آید.

در این سیستم ابتدا با کمک فایل result.htm که نتایج حاصل از موتور جستجو رالیست می کند صفحات لیست می شود سپس کاربر صفحاتی را که مورد علاقه اش است انتخاب می کند که به آن hit می گویند و صفحاتی که انتخاب نمی کند missگفته می شود سپس در یک rerank صفحات hit و صفحاتی که بیشتر مشابه این صفحات هستند لیست می شوند.به این صورت سعی می شود نتایج نزدیک به علاقه و نیاز کاربر باشد.همانطور که دیده می شود در این روش نیز کاربر دخالت صریح دارد.

در [29] روند کار به این صورت است که ابتدا همه صفحات بازیابي شده در کلاس بد قرار میگیرند. پس از اینکه کاربر بر روي يكي از نتايج موجود در ليست بردا كليک نمود، با اين فرض كه صفحه كليک شده مورد علاقه كاربر مي باشد، آن را به كلاس خوب منتقل مينمايد و با استفاده از طبقه بندي كننده Naive Bayesian يک نمره به هر صفحه در ليست بردا اختصاص ميدهد. سپس ليست بر اساس نمرات اختصاص يافته رتبه بندي شده و به كاربر ارائه مي شود تا كاربر صفحه ديگري را به عنوان صفحه مورد علاقه انتخاب نمايد و مراحل قبلي مجددا تكرار مي شود.

Excalibur سمت سرور كار مي كند و پروفايل كاربر را به صورت يک كوكي روي سيستم او ذخيره مي كند .در اين سيستم نيز چون كاربر بايد در حين جستجو هر بار صفحات مورد علاقه خود را انتخاب نمايد و سيستم چندين بار با توجه به علاقه كاربر rerank می شود حالت ضمنی وجود ندارد.

کاربران معمولا به انتشار علایقشان در اینترنت بد بین هستند و همچنین این کار برایشان وقت گیر و هزینه بر است و ممکن است که در بیان علایقشان دچار اشتباه شوند یا به خاطر عدم اعتمادی که به فضای وب

دارند اطلاعات صحیحی را به سیستم ندهند, مجموع این دلایل باعث می شود که جمع آوری پروفایل کاربر به صورت اتومات نتایج بهتری را به دست آورد.

2-2-مدل سازی ضمنی کاربر

در رویکرد دوم سعی شده که دخالت کاربر به حداقل برسد و پروفایل او به صورت ضمنی از روی تعاملاتی که سیستم با کاربر دارد مثل مرورهای کاربر،پرس وجوهای گذشته او ،ایمیلهایی که مشاهده کرده و... شناخته میشود.مثل کارهای انجام شده در [30],[10],[16],[17].

در [30] سعي شده که کمترین دخالت کاربر وجود داشته باشد.آنها پروفایل کاربران را بر اساس سابقه جستجوي کاربران در موتور جستجو مي سازند.در آن کار یک لفافه بند براي گوگل (google wrapper) اجرا شده است.

منابع مختلف پروفایل، پرس و جو ها و نتایج جستجو ها است.این پروفایل با کلاسبندي اطلاعات در محتویات پروژه دایرکتوري باز ارثي ساخته مي شود و سپس براي دوباره مرتب کردن نتایج جستجو استفاده مي شود از بازخورد هاي کاربر استفاده مي شود تا ترتیب نتایج گوگل با ترتیب جدید ما مقایسه شود وتا حد 37 نتایج بهبود یافته است.پایه این کار بر ساختن پروفایل کاربر از تعاملات کاربر با یک موتور جستجوي خاص است.در این کار از google wrapper استفاده شده است یعني یک لفافه حول موتور جستجوي گوگل براي ثبت کردن پرس وجو ها و نتایج جستجوها وکلیکها بر پایه هر کاربر قرار می گیرد.ایراد این کار این است که اولا فقط یک موتور جستجوی خاص را بررسی می کند و ثانیا سمت سرور است و به غیر از جستجوهای کاربر از وبگردی ها ورفتار کلی او در وب اطلاعاتی نخواهیم داشت .

3- کارهای انجام شده از منظر نحوه اجرای کار

از نظر نحوه انجام کار نیز دو رویکرد وجود دارد .

3-1- روش قطعی

در رویکرد اول با این مساله به صورت یک مساله crisp بر خورد شده است.در این سیستمها پس از ساختن پروفایل کاربر ، کاربران با توجه به پروفایل شان به صورت قطعی و همیشگی عضو یک گروه می شوند و صفحات نیز پس از بررسی محتویاتشان به صورت قطعی گروهبندی می شوند و ارتباط بین مفاهیم و کاربران نیز به صورت قطعی بر قرار می شود .مثل کارهای انجام شده در [3],[4],[16][31],,[17]

در [31] از الگوریتم BM25 برای مرتب کردن صفحات با توجه به پرس وجوی ارائه شده به سیستم استفاده می شود.در این روش با درنظرگرفتن یک مجموعه N تایی شامل کل اسناد موجود در وب و یک مجموعه R تایی شامل اسناد مورد علاقه کاربر، وزن هر مفهوم با استفاده از فرمول زیر محاسبه می شود:

در فرمول بالا ri تعداد اسناد موجود در وب شامل مفهوم i ام و ni تعداد اسناد مورد علاقه کاربر شامل مفهوم i ام می باشد.

در [30] کلاسبندي کننده بر اساس کلیدهاي محتوایي که از الگوریتم k نزدیکترین همسایه استفاده مي کند.یک مجموعه از اسکریپتها که log file را پردازش مي کند و کارایي آن را براي هر کاربر ارزیابي مي کند استفاده مي شود که آن را به مجموعه تست و آموزش تقسیم مي کند. پروفایل کاربر به صورت یک سلسله مراتب وزندار نمایش داده مي شود. در این سیستم یک پروفایل کاربر داریم یک پروفایل اسناد که شباهت

مفهومي بين آنها با فرمول زير محاسبه مي شود:

2-3- روش فازی

اما در رویکرد دوم با این مساله با دید فازی برخورد شده است.عدم قطعیت این مسالله در پیدا کردن مفاهیم اسناد و هم در تشخیص نیاز کاربر از پرس وجوی او و هم در ارتباط بین سند و مفهوم مورد نیاز کاربر وجود دارد و همچنین ارتباط بین پرس و جو ها و مفاهیم نیز دارای سطحی از عدم قطعیت است.ما نمی توانیم ادعا کنیم که کاربری به صورت قطعی عضو یک گروه است و به گروههای دیگر هیچ علاقه ای ندارد.این خوشه بندی ماهیت غیر قطعی وفازی دارد و برخورد قطعی با آن نا مناسب است.
سیستمهایی که از روش فازی برای دسته بندی کاربران و محتویات صفحات و مفاهیم استفاده کنند این عدم قطعیت را در نظر گرفته اند.مثل کار های انجام شده در [9],[10],[1]

در [9] شبکه مفهومی فازی مسئول شخصی سازی لینکهای مرتبط به پرس و جوی کاربر است . این شبکه را می توان برای هرکاربر با توجه به پروفایل او ساخت. پس از ارسال پرس وجوي كاربر به يک موتور جستجوي متني، نتايج بازيابی شده در مجموعه ريشه قرار ميگيرند، سپس صفحات وبی که به صفحات موجود در مجموعه ريشه لينک دارند و صفحات وبی که اين صفحات به آنها لينک دارند نيز به مجموعه ريشه اضافه می شوند .در مجموعه حاصل صفحاتی که بيشترين لينک به آنها وجود دارد، يعنی صفحات معتبر، و صفحاتی که بيشترين لينک را به صفحات ديگر دارند، يعنی صفحات هاب، تعيين می شوند که الگوريتم هايی برای تعيين آنها نيز معرفی می شود .حال از رويکرد شبکه مفهومی فازي برای رتبه بندي صفحات معتبر به دست آمده بر اساس علاقه کاربر استفاده می شود . در واقع استفاده از ساختار پيوند علاوه بر شبکه مفهومی فازي، شخصی سازي نتايج جستجو بر اساس علايق کاربران را بهبود می دهد. سيستم در يک شبکه مفهومی فازی مفاهيم و صفحات مرتبط به آنها را قرار می دهد.سپس نياز کاربر را از پروفايل او استخراج می کند و سپس صفحات مرتبط با نياز کاربر را با توجه به ساختار لينک و مفاهيم مورد علاقه کاربرپيدا کرده و در ليست مرتب شده قرار می دهد.

اما اين پروفايل با دخالت کاربر ساخته می شود و همچنين شبکه با نظارت فرد خبره ساخته می شود که امر مطلوبی نيست.

در[1] شخصی سازي نتايج موتور جستجو با استفاده از شبکه مفهومی فازي گسترش يافته انجام می شود .در اين کار پروفايل کاربر و نتايج جستجو با يک شبکه مفهومی فازي گسترش يافته نمايش داده می شوند .در شبکه مفهومی فازي گسترش يافته 4 نوع ارتباط فازي که شامل ارتباط فازي مثبت، ارتباط فازي منفی، ارتباط فازي عمومی و ارتباط فازي خصوصی می باشد، می تواند بين مفاهيم وجود داشته باشد که با يک ماتريس رابطه نمايش داده ميشود. اين شبکه نيز توسط فرد خبره ساخته می شود.

همانطور که ديده می شود کارهای انجام شده در گذشته يا پروفايل کاربر را به صورت صريح می سازند و يا اگر پروفايل به روش صريح ايجاد شود در اجرای کار از روشهای crisp استفاده می کنند و سيستمی وجود ندارد که در آن هم پروفايل کاربر به صورت ضمنی واتوماتيک ساخته شود و هم از روشهای فازی استفاده شده باشد.
البته در [9],[1] تا حدودی سعی شده است که شبکه مفهومی فازی با پروفايل اتوماتيک راه اندازی شود اما ايراد اين سيستم هم اين است که در آن هر چند طبقه بندی مفاهيم

فازی است اما گروهبندی کاربران به صورت crisp است که با عدم قطعیت مساله همخوانی ندارد.
در مجموع در هیچ یک از کارهای گذشته تهیه پروفایل کاربر به صورت اتوماتیک و ضمنی و به روز رسانی پروفایل او و تولید شبکه فازی به صورت اتوماتیک و به روز رسانی آن و کلاسبندی فازی کاربران با هم انجام نشده است و در بعضی مقالات از روشهای فازی استفاده شده اما شبکه فازی توسط فرد خبره تولید می شود و این کار اتوماتیک نیست.فقط در یک مقاله این کار اتوماتیک انجام شده که در آن به روز رسانی شبکه فازی و به روز رسانی پروفایلها وجود ندارد.همچنین کلاسبندی فازی کاربران در هیچ کدام از کارهای قلبلی وجود ندارد.در همه کارهای انجام شده حتی کارهایی که از روشهای فازی استفاده می کنند کلاسبندی به صورت crisp است که این امر اصل فازی بودن و غیر قطعی بودن موضوع را زیر سوال می برد.همچنین در هیچ کدام از کارهای انجام شده به روز رسانی شبکه فازی با گذشت زمان و با تغییر پروفایل کاربر وجود ندارد که این امر بسیار مهمی است،چون علایق و دانش کاربران در طول زمان تغییر می کند.

مراجع

[1] S. M. Chen, Y. J. Horng and C. H. Lee, "Fuzzy Information Retrieval based onMulti-relationship Fuzzy Concept Networks", Fuzzy Sets and Systems, vol. 140, pp.183-205, 2003

[2] P. Ferragina and A. Gullì, "A Personalized Search Engine Based on Web Snippet Hierarchical Clustering", Proceedings of the World Wide Web Conference (WWW) (The Tokio, The Japan), pp. 801-810, 2005

[3] W. Alhalabi, M. Kubat and M. Tapia, "Search Engine Personalization Tool Using Linear Vector Algorithm", Proceedings of the 4th Saudi Technical Conference andExhibition, pp. 336-344, 2006

[4] M. Radovanovic and M. Ivanovic, "CatS: A Classification-Powered Meta-Search Engine", Advances in Web Intelligence and Data Mining, Springer-Verlag, vol. 23,pp. 191–200, 2006

[5] W. Kim, L. Kerschberg, A.Scime,"Learning for automatic personalization in a semantic taxonomybased meta-search agent",Elsevier,Electronic Commerce Research and Applications 1 (2002) 150–173

[6] K . J . Kim, S . B . Cho," Personalized mining of web documents using link structures and fuzzy concept networks",Elsevier, Applied Soft Computing 7 (2007) 398–410,2005

[7] Hong Zhang1, Yanhong Ma2, Qiuyu Zhang1, Pengshou Xie1, Zhongxian Bao "Personalized Intelligent Search Engine Basedon Web Data Mining"Proceedings of the 2009 International Workshop on Information Security and Application (IWISA 2009)Qingdao, China, November 21-22, 2009

[8] Mariam Daoud , Lynda-Tamine Lechani,Mohand Boughanem,"Towards a graph-based user profile modeling f or a session-based personalized search ",Springer-Verlag London Limited 2009

[9]C . H . Lee,Y . H . Kim,P . K . Rhee,"web personalization expert with combining collaborative filtering and association rule mining technique",Elsevier,expert system with application 21(2001) 131-137

[10]P . Kazienko , M . Adamski,"Adrosa – adaptive personalization of web advertising",Elsevier,information sciences 177(2007) 2269-2295

[11] B . P . C . Yen, R . C . W. Kong, "Personalization of information access for electronic catalogs on the web",Elsevier,Electronic commerceresearch and applications 1 (2002) 20 – 40

[12]Y . F . kuo,L . S . Chen,"personalization technology application to internet content provider", Elsevier, Expert Systems with Applications 21 (2001) 203 – 215

[13] M. Clements , A.P. de Vries , M.J.T. Reinders," The influence of personalization on tag query length in social media search",Elsevier, Information Processing and Management 46 (2010) 403–412

[14] L. Yuen, M. Chang, Y. K. Lai and C. K. Poon, "Excalibur: a Personalized Metasearch

Engine", Proc. 28th Annual International Computer Software and Applications Conference (COMPSAC'04), vol. 2, pp.49-50, 2004.
[15] M. Speretta and S. Gauch, "Personalizing Search Based on User Search History",Submitted to CIKM ‘04, (2004).
[16] J. Teevan, S. T. Dumais and E. Horvitz, "Personalizing Search via Automated Analysis of Interests and Activities", Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 05), ACM Press, pp.449-456, 2005.
[17] S. Souldatos, T. Dalamagas and T. Sellis, "Captain Nemo: A Meta-Search Enginewith Personalized Hierarchical Search Space", INFORMATICA -LJUBLJANA- ,vol. 30, pp. 173-182, 2006