امیرحسین کیهانی‌پور کارشناسی خود را در رشته مهندسی کامپیوتر-نرم‌افزار از دانشگاه صنعتی شریف و کارشناسی ارشد خود را در رشته مهندسی کامپیوتر-هوش ماشین و رباتیک از دانشگاه تهران دریافت نموده است. او هم اکنون مشغول به تحصیل در مقطع دکتری رشته مهندسی کامپیوتر-هوش ماشین و رباتیک در دانشگاه تهران می‌باشد. فعالیتهای پژوهشی او در زمینه بازیابی اطلاعات وب، نظریه ترکیب اطلاعات و نیز الگوریتمهای یادگیری ماشین می‌باشد.

واژه نامه:

[1] Adaptive
[2] Term Frequency
[3] In-Degree
[4] Ordered Weighted Aggregation Operator
[5] Benchmark
[6] http://research.microsoft.com/users/tyliu/LETOR/
[7] Ranking
[8] Content
[9] Connectivity
[10] Boolean
[11] Vector Space
[12] Rank spamming
[13] Web Information Retrieval
[14] Links
[15] Precision
[16] Ordered Weighted Operator
[17] Output links
[18] Sinking pages
[19] Damping Factor
[20] Host
[21] Domain
[22] Super-node
[23] Term Frequency
[24] Inverse Document Frequency
[25] Okapi
[26] Relevance propagation
[27] Score
[28] Hyper-relevance
[29] Sitemap-based term propagation
[30] Aggregation
[31] n-tuple
[32] Constrained
[33] Feature
[34] Fine-grained
[35] Coarse-grained
[36] Popular
[37] Rich-get-richer
[38] Error Back Propagation
[39] Average Precision
[40] Mean Average Precision
[41] Borda Voting
[42] Off-line
[43] Combination Rank Fine-Grained
[44] Document Length

[2] McBryan, O.A., GENVL and WWWW: Tools for Taming the Web,In Proceedings of the First International World Wide Web Conference, pp. 79-90, 1994.
[3] Gulli, A., & Signorini, A., "The Indexable web is more than 11.5 billion pages", In Proceedings of the 14th international conference on World Wide Web, ACM Press, pp. 902-903, 2005.
[4] http://www.worldwidewebsize.com/, Jan. 2008.
[5] http://www.internetnews.com/stats/article.php/1363881
[6] http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf,2005.
[7] Baeza-Yates, R., & Ribeiro-Neto, B., Modern Information Retrieval. ACM Press/ Addison-Wesley, 1999.
[8] Salton, G., & Buckley, C. Term weighting approaches in automatic text retrieval. Information Processing and Management, Volume 24, Issue 5, pp. 513–523, 1988.
[9] Robertson, S. E., Walker, S., HancockBeaulieu, M. M., Gatford, M., & Payne, A. Okapi at TREC-4. In NIST Special Publication. The Fourth Text REtrieval Conference (TREC-4), pp. 7396, 1995.
[10] Henzinger, M., Motwani, R., & Silverstein, C., Challenges in web search engines, SIGIR Forum, Volume 36 , Issue 2, pp. 11-22, 2002.
[11] Henzinger, M., Hyperlink analysis for the web. IEEE Internet Computing, Volume 5, Issue 1, pp. 45–50, 2001.
[12] Page, L., Brin, S., Motwani, R., & Winograd, T. The PageRank citation ranking: Bringing order to the web. Technical report, Computer Science Department, Stanford University, 1998.
[13] Kleinberg, J. M. Authoritative sources in a hyperlinked environment. Journal of the ACM, Volume 46, Issue 5, pp.604–632, 1999.
[14] Zareh Bidoki, A.M. and Yazdani, N., DistanceRank: An Intelligent Ranking Algorithm for Web Pages, Information Processing & Management, Volume 44, Issue 2, pp. 877-892, 2008.
[15] Liu, T. Y., Xu, J., Qin, T., Xiong, W., and Li, H. LETOR: Benchmarking learning to rank for information retrieval. In Proceedings of SIGIR Workshop on Learning to Rank for Information Retrieval , 2007.
[16] Najork, M., Zaragoza, H., & Taylor, M. J., Hits on the web: how does it compare? In Proceedings of SIGIR, pp. 471-478, 2007.
[17] Yager, R.R., On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Transactions on Systems, Man and Cybernetics, Volume 18, Issue 1, 1988.
[18] TREC, http://trec.nist.gov/data/webmain.html, 2004.
[19] Robertson, S.E., Walker, S., Microsoft Cambridge at TREC-9: Filtering track, *The Ninth Text REtrieval Conference (TREC-9)*, pp. 73-86, 2000.
[20] O'Hagan,M., Aggregating template rule antecedents in real-time expert systems with fuzzy set logic, In Proceedings of 22th Annual IEEE Asilomar Conference on Signals, Systems and Computers, pp. 681-689, 1988.
[21] Filev , D. & Yager, R. R., On the issue of obtaining OWA operator weights, Journal of Fuzzy Sets and Systems, Volume. 94, Issue. 2, pp. 157-169, 1998.
[22] Cho, J., Roy, S., & E. Adams, R., Page Quality: In Search of an Unbiased Web Ranking. In Proceedings of ACM International Conference on Management of Data (SIGMOD), 2005.
[23] Borda, J. C. Memoire sur les elections au scrutin. Histoire de l'Academie Royale des Sciences, 1781.
[24] Filev, D., and Yager, R. R., Learning OWA operator weights from data. Proceedings of the third IEEE Conference on Fuzzy Systems, Volume. 1, pp. 468-473, 1994.
[25] LETOR,http://research.microsoft.com/users/tyliu/LETOR 2007.
[26] Jarvelin, K., & Kekalainen, J., IR evaluation methods for retrieving highly relevant documents. In Proceedings of the ACM Conference on Research and Development on Information Retrieval (SIGIR), 2000.
[27] Xue, G. R., Yang, Q., Zeng, H. J., Yu, Y., and Chen, Z., Exploiting the hierarchical structure for link analysis, SIGIR Conference, 2005.
[28] Qin, T., Liu, T.-Y., Zhang, X.-D., Chen, Z., & Ma, W.-Y., A study of relevance propagation for web search. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 408–415, 2005.
[29] Shakery, A., Zhai, C. X. Relevance Propagation for Topic Distillation UIUC TREC 2003 Web Track Experiments, Proceedings of TREC, 2003.

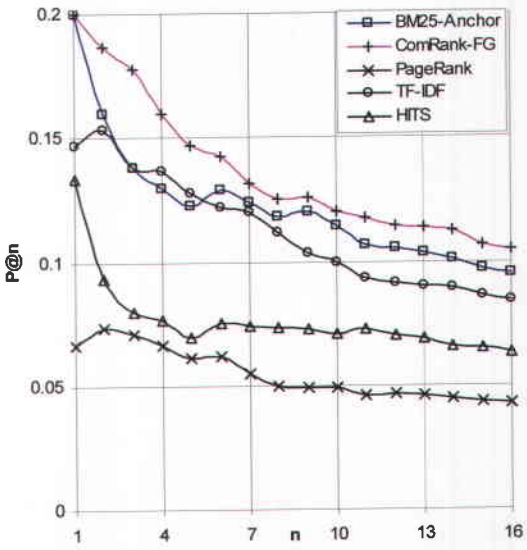
علی‌محمد زارع‌بیدکی کارشناسی خود را در رشته مهندسی کامپیوتر- سخت‌افزار از دانشگاه صنعتی اصفهان و کارشناسی ارشد خود را در همین رشته از دانشگاه تهران دریافت نموده است. او همچنین دکتری خود را در رشته مهندسی کامپیوتر-نرم‌افزار در سال ۱۳۸۸ از دانشگاه تهران دریافت نموده است. او هم‌اکنون استادیار دانشکده مهندسی برق و کامپیوتر دانشگاه یزد می‌باشد. فعالیتهای پژوهشی او در زمینه بازیابی اطلاعات وب، پردازش زبان‌های طبیعی و شبکه‌های کامپیوتری می باشد.

محمد آزادنیا کارشناسی خود را در رشته مهندسی مخابرات از دانشگاه علم و صنعت و کارشناسی ارشد خود را در رشته مهندسی صنایع از دانشگاه صنعتی شریف دریافت نموده است. او از سال ۱۹۶۷ بعنوان پژوهشگر و مدیر پروژه و از سال ۱۳۷۹ به عنوان عضو هیئت علمی در مرکز تحقیقات مخابرات ایران فعالیت داشته است. فعالیتهای پژوهشی او بیشتر در زمینه‌های سیستمهای اطلاعاتی، مدیریت فناوری اطلاعات، پردازش زبان طبیعی و بازیابی اطلاعات وب می باشد.

ناصر یزدانی کارشناسی خود را در رشته مهندسی کامپیوتر از دانشگاه صنعتی شریف و دکتری خود را در همین رشته از دانشگاه *کیس وسترن رزرو* در آمریکا دریافت نموده است. او همچنین در شرکت‌ها و موسسه‌های تحقیقاتی مختلف آمریکا و ایران از جمله مرکز تحقیقات مخابرات ایران فعالیت داشته است. او هم‌اکنون دانشیار دانشکده مهندسی برق و کامپیوتر دانشگاه تهران می‌باشد. فعالیتهای پژوهشی او در زمینه شبکه‌های کامپیوتری، بازیابی اطلاعات وب، سیستم‌های عامل و پایگاه‌های داده می باشد.

ترکیب شده‌اند. لازم به ذکر است که هدف استفاده از In-Degree (درجه ورودی) صفحات به عنوان یک ویژگی می‌باشد، ولی با توجه در دسترس نبودن گراف، از این دو الگوریتم که به صورت برون‌خط با استفاده از درجه ورودی‌ها محاسبه شده‌اند، استفاده شده است. بنابراین هر سه ویژگی استفاده شده‌ی فوق ساده می‌باشند.

الگوریتم ترکیبی حاصل ComRank-FG نامیده شده و مانند روش ارائه شده در بخش ۴ عمل می‌نماید و نظر کاربران به عنوان تابع هدف در نظر گرفته شده است. بنابراین مانند بخش قبل فرآیند ارزیابی شامل دو بخش آموزش (۵۰ پرسش شامل ۴۰ هزار جفت پرسش و سند) و آزمون (۲۵ پرسش) می باشد.



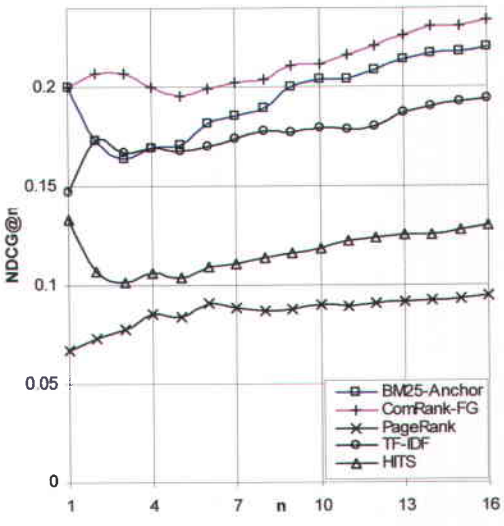
شکل (۸): مقایسه الگوریتم ComRank-FG با الگوریتم‌های دیگر در P@n

برای ارزیابی این الگوریتم با توجه به اینکه تمام ویژگیهای الگوریتم BM25 برای متن در دست نیست از ویژگیهای متن Anchor ها مانند TF استفاده شده است. همچنین با BM25 حاصل از متن Anchor عمل مقایسه انجام گرفته است.

در شکل‌های ۸ و ۹ الگوریتم ComaRank-FG با بقیه الگوریتم‌ها مقایسه شده‌اند. همانطور که نشان داده شده است، الگوریتم ارائه شده از بقیه بهتر عمل کرده است. این در حالیست که دارای پیچیدگی پایینی نسبت به BM25 می‌باشد و وزن‌ها به صورت پویا محاسبه می‌شوند. علت اصلی این موفقیت ترکیب محتوا و ساختار وب به صورت مناسب بوده است. به عبارت دیگر با ترکیب ویژگیهای محتوا و ساختار علاوه بر حل مشکلات موجود، جواب بهتری نیز ارائه گردیده است. بدین ترتیب ادعای فوق مبنی بر حصول یک الگوریتم دلخواه از طریق استفاده از ویژگی‌های ریزدانه ثابت می‌گردد.

۶- نتیجه گیری و کارهای آینده

مهمترین بخش یک موتور جستجو را واحد رتبه بندی تشکیل می‌دهد. در رتبه بندی، نتایج به ترتیب درجه ارتباط آنها با پرسش، مرتب شده و به کاربر ارائه می‌شوند. به عبارت دیگر هدف اصلی در رتبه‌بندی صفحات کشف کیفیت واقعی آنها می‌باشد. در حال حاضر روشهای مختلفی مبتنی بر اتصال مانند PageRank و مبتنی بر محتوا مانند TF-IDF و BM25 ارائه شده اند.



شکل (۹): مقایسه ComRank-FG با الگوریتم‌های دیگر در NDCG@n

در این مقاله دو روش ترکیبیِ وفقی به نامهای ComRank و ComRank-FG ارائه شده‌اند در مورد روش اول چندین الگوریتم رتبه بندی مشهور به عنوان ویژگی‌های درشت‌دانه و در روش دوم چندین ویژگی ریزدانه مانند TF و In-Degree با یکدیگر ترکیب شده‌اند. ترکیب روشها به دو صورت استفاده از رأی گیری وزن‌دار و عملگر تجمیع OWA انجام می‌گیرد. در روش رأی‌گیری وزن هر ویژگی (الگورینم) متناظر با دقت آن تعیین می‌گردند. همچنین ویژگیها با استفاده از عمگر تجمیع OWA نیز ترکیب شده‌اند که در آن وزن‌ها به صورت پویا با استفاده از نظر افراد در رابطه با درجه‌ی ارتباط یک سند با پرسش یادگرفته می‌شوند.

برای ارزیابی و تست روش ارائه شده از بسته LETOR شامل مجموعه داده‌های محک کنفرانس TREC 2004 استفاده شده است. نتایج بدست آمده بهبود چمشگیری را در روش ترکیبی ارائه شده در مقایسه با بقیه الگوریتم‌ها نشان می‌دهد.

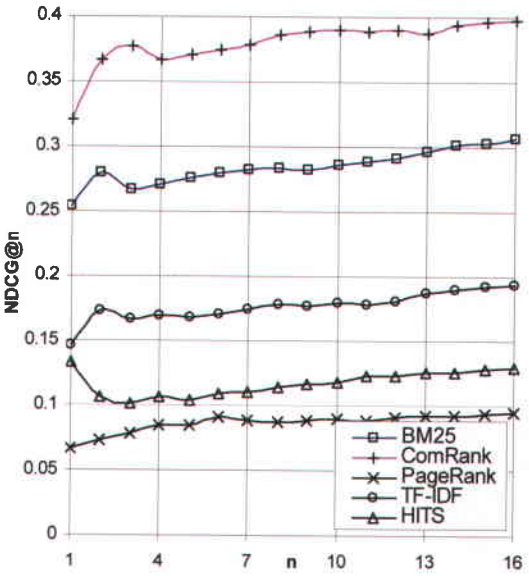
گرچه روش ترکیب‌کننده الگوریتم های مختلف دارای دقت بالایی می‌باشد ولی پیچیدگی آن نسبتاً زیاد است. در حالیکه الگوریتم ترکیبی ارائه شده‌ی مبتنی بر ویژگیهای ریزدانه علاوه بر داشتن دقت مناسب دارای پیچیدگی پایینی است. علاوه بر آن روشهای ارائه شده به علت داشتن دقت بالا در چند نتیجه اول، برای محیط وب بسیار مناسب می باشند. نتایج فوق نشان دهنده اهمیت ترکیب ویژگی‌های محتوایی و ساختاری در روشهای رتبه بندی است. ارائه یک الگوریتم مناسب رتبه بندی با استفاده از شبکه های عصبی و الگوریتم‌های ژنتیک بر اساس ویژگیهای ریزدانه، همچنین پیاده سازی الگوریتم رتبه‌بندی فوق در موتور جستجوی فارسی به عنوان کارهای آینده پیشنهاد می‌گردند.

مراجع

[۱] زارع‌بیدکی ع. م.،آزادنیا م.، یزدانی ن. و کیهانی‌پور ا.ح. "الگوریتم ترکیبیِ وفقی جهت رتبه‌بندی صفحات وب"، چاپ شده در سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران، اسفند ۱۳۸۶.

با تفاسیر فوق الگوریتم ترکیبی ارائه شده (ComRank) با الگوریتم‌های PageRank و BM25 مقایسه شده است. شکلهای (۶) و (۷) به ترتیب مقایسه الگوریتم ComRank در مقایسه با دو الگوریتم BM25 و PageRank را برای دو معیار P@n و NDCG را نشان می‌دهد.

همانطور که در شکلها نشان داده شده است الگوریتم‌های ارائه شده در مقایسه با روشهای موجود کارکرد بهتری دارند. خصوصاً در n های کوچک مقدار آن حدود ۱۰٪ بهتر از BM25 خصوصاً برای فاکتور NDCG می‌باشد.



شکل (۷): مقایسه ComRank با الگوریتم‌های دیگر در NDCG@n

با توجه به اینکه کاربرهای وب معمولاً جوابهای اول را کلیک می‌کنند، الگوریتم فوق برای محیط وب بسیار مناسب خواهد بود.

مطابق شکل کارآیی الگوریتم ComRank حدود ۴٪ کمتر از الگوریتم بوردا است. با توجه به اینکه روش بوردا به عنوان تابع هدف تعیین شده است، نتیجه فوق قابل قبول می‌باشد. بنابراین با داشتن یک مجموعه محک که برای هر جفت پرسش یک عدد بین صفر و یک را داشته باشد می‌توان به جواب‌های بهتری نیز دست یافت.

بنابراین از بحثهای فوق نتیجه می‌شود که از ترکیب چندین الگوریتم رتبه بندی می‌توان به یک روش جدید با کارآیی بالایی دست پیدا کرد. جدول ۴ میانگین دقت (MAP) را برای روشهای مختلف ارائه شده نشان می‌دهد.

جدول (۴): میانگین دقت (MAP) الگوریتم‌های رتبه بندی.

نام الگوریتم	میانگین دقت برای همه پرسشها
Borda	۰/۲۶
BornaNorm	۰/۳۳
ComRank	۰/۳۰
BM25	۰/۲۴
PageRank	۰/۰۶

۵-۱- پیچیدگی الگوریتم ارائه شده

پیچیدگی الگوریتم ترکیبی ارائه شده را می‌توان از دو جنبه محاسبه‌ی وزن‌ها و استفاده از الگوریتم در یک سیستم جستجو بررسی کرد.

پیچیدگی محاسباتی الگوریتم برابر O(mnr) می‌باشد که m، n و r به ترتیب نشانگر تعداد جفت‌های سند و پرسش، تعداد ویژگی‌ها (۵ ویژگی) و تعداد تکرار هستند. مطابق آزمایشات انجام شده برای دقت ۰/۰۰۱ ده تکرار کافی می‌باشد.[1] اگر چه محاسبه وزن‌ها به صورت بازگشتی انجام می‌شود ولیکن یک فرایند برون خط[42] می‌باشد و در نهایت اثری روی زمان پاسخ به کاربران نخواهد داشت. به عبارت دیگر در ابتدا وزن‌ها با استفاده از داده‌های آموزش محاسبه شده و در سیستم جستجو با استفاده از وزن‌های محاسبه شده از قبل عمل ترکیب صورت می‌گیرد. جنبه‌ی دوم ترکیب چند الگوریتم مختلف با یکدیگر می‌باشد که در سیستم‌ها و موتورهای جستجوی مختلف نیز استفاده شده است[25]. گرچه در ظاهر پیچیده به نظر می‌رسد ولیکن با توجه به اینکه اکثر اعضای ترکیب از قبل محاسبه شده اند پیچیدگی این روش کمتر خواهد بود. برای مثال در روش ComRank که شامل ترکیب ۵ روش مختلف می‌باشد، الگوریتم‌های HostRank از قبل به صورت برون خط محاسبه شده‌است و همچنین اجزای بقیه‌ی روش‌ها که شامل TF، IDF و درجه وروی می‌باشد نیز از قبل وجود دارند. بنابراین روش فوق به راحتی همانطور که در بقیه‌ی سیستم ها استفاده شده است قابل استفاده خواهد بود. همچنین روش ارائه شده در بخش بعد در مقایسه با ComRank پیچیدگی بسیار کمی دارد.

۵-۲- الگوریتم ترکیبی ریز دانه (ComRank-FG[43])

اگرچه با ترکیب روشهای مختلف رتبه بندی در بخش قبلی به نتایج مطلوبی رسیده‌ایم ولیکن این روش‌ها مشکلاتی هم دارند. مهمترین مشکل روش قبل پیچیده‌گی آن می‌باشد. به عبارت دیگری لازم است چندین روش رتبه بندی برای رسیدن به یک روش محاسبه شوند. علت اصلی آن درشت‌دانه بودن ویژگیهای استفاده شده است. بنابراین هدف پیدا کردن ترکیبی از ویژگیهای ریزدانه برای رتبه بندی می‌باشد تا علاوه بر رتبه بندی مناسب دارای پیچیدگی کمتری باشند. بدین منظور از ویژگیهای ریزدانه را می‌توان فرکانس کلمه(TF)، (درجه‌ی ورودی یک صفحه‌ی وب) In-Degree و (طول سند) DL[44] نام برد. در اینجا سعی داریم ترکیب مناسبی از آنها را ارائه نماییم.

هر تابع پیچیده را می‌توان به صورت یک تابع دو جمله‌ای نوشت. برای مثل قسمت اصلی الگوریتم BM25 (بخش ۲)، عبارت است از $\frac{tf}{k+tf}$ که با استفاده از سری تیلور میتوان آن را به صورت زیر بیان نمود:

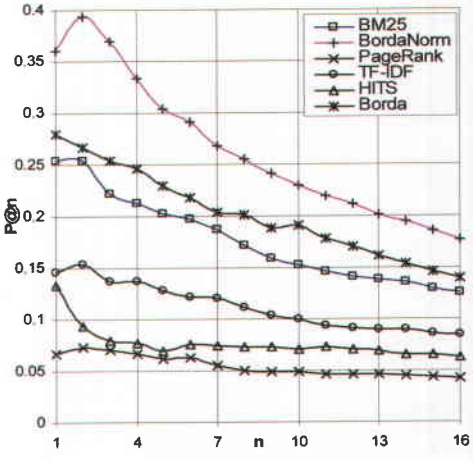
$$f(x)=\frac{x}{k+x}=\sum_{i=1}^{\infty}(-1)^{i+1}(\frac{x}{k})^{i}=$$
$$x-\frac{x^2}{k^2}+\frac{x^3}{k^3}-\frac{x^4}{k^4}+\ldots$$

به عبارت دیگر می‌توان با استفاده از ویژگیهای ریزدانه به یک الگوریتم رتبه بندی دلخواه با کارآیی مناسب دست یافت.

در این بخش سه ویژگی TF، PageRank و HostRank با استفاده از الگوریتم ارائه شده در شکل ۳ (که مبتنی بر OWA است) با یکدیگر

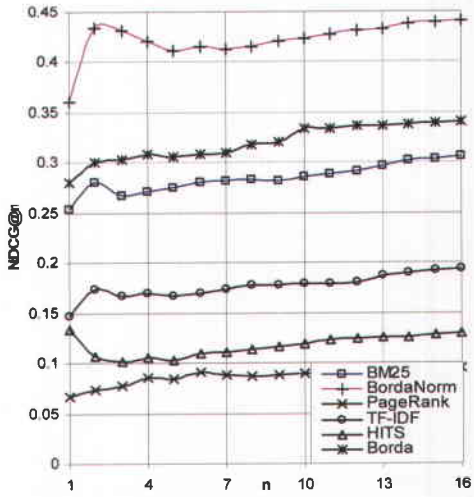
[1] زمان اجرا روی یک PC با یک گیگابایت حافظه کمتر از سه دقیقه به طول انجامید.

در شکلهای (۴) و(۵) الگوریتم ترکیبی بوردا در دو حالت بدون نرمال سازی و با نرمال سازی با بقیه روشها در دو معیار P@n و NDCG مقایسه گردیده‌اند.



شکل (۴): مقایسه الگوریتم بوردا با الگوریتم‌های دیگر در P@n در دوحالت نرمال‌سازی و بدون نرمال‌سازی

همانطور که دیده می‌شود روش رأی گیری بوردا که در آن الگوریتم‌های مختلف با استفاده از فرمول (۱۷) نرمال شده‌اند نسبت به بقیه افزایش چشمگیری داشته است. برای مثال در اولین نتیجه (n=1) و دومین نتیجه (n=2) به ترتیب حدود ۱۱٪ و ۱۵٪ نسبت به BM25 بهتر عمل کرده‌اند و نسبت به نتایج بدست آمده در مقاله قبلی [1] حدود ۵٪ بهتر است. همانطور که ذکر شد، علت آن همگن (نرمال) شدن آرگومانهای الگوریتم‌های استفاده می‌باشد. نکته‌ی دیگری که در شکل‌ها نمایان است افزایشی بودن معیار P@n و (تقریباً) ثابت ماندن معیار NDCG با افزایش n است. شکل نمودار به صورت کلی به مجموعه‌ی داده‌ها و الگوریتم استفاده شده بستگی دارد. ولیکن در یک الگوریتم مؤثر با افزایش تعداد جواب‌ها (n) بایستی تعداد نتایج مرتبط کم شده و در نتیجه معیار دقت کاهشی و معیار NDCG (تقریباً) ثابت خواهند بود. با توجه به اینکه کاربران اغلب نتایج اول (۱۰ نتیجه) یک موتور جستجو را کلیک می‌کنند بنابراین هدف الگوریتم نیز داشتن دقت بالا در نتایج اولیه می‌باشد.



شکل (۵): مقایسه الگوریتم بوردا با الگوریتم‌های دیگر در NDCG@n در دوحالت نرمال‌سازی شده و بدون نرمال‌سازی

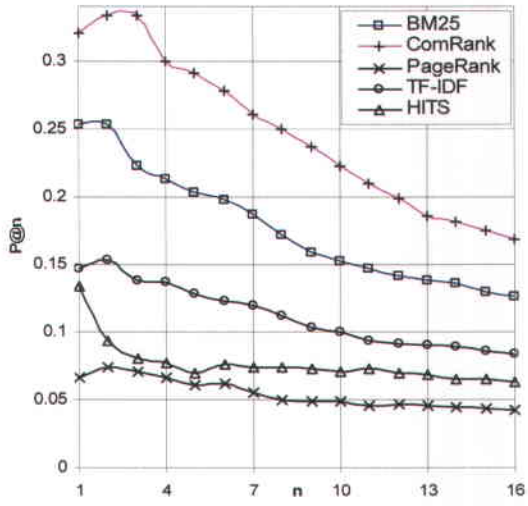
با توجه به اینکه محتوای وب داری پویایی بالایی می‌باشد، مهمترین مشکلی که روش فوق دارد ثابت بودن ضرایب ویژگیها (الگوریتم‌های رتبه بندی) می‌باشد (جدول ۳). بنابراین به ارزیابی روش ترکیبی ارائه شده در شکل (۳) که مبتنی بر یادگیری است و ورن ویژگی‌ها متغیر می باشند، می‌پردازیم.

مهمترین مسئله در یادگیری داشتن تابع هدف جهت پیدا کردن وزن‌های مناسب است. بهترین تابع درجه‌ی مرتبط بودن پرسش با سند از دید کاربر می‌باشد. با این تابع هدف با دو مشکل عمده مواجه می‌شویم. اولین مشکل از هم نوع نبودن درجه‌ی ارتباط با مقدار حاصل از ترکیب ویژگیها که همان الگوریتم‌های رتبه بندی هستند، به وجود می‌آید. به علاوه چونکه در مجموعه‌ی داده‌های موجود فقط دو سطح ۰ و ۱ توسط کاربر ارائه شده است تابع یادگیری با مشکل مواجه خواهد شد (جدول(۲)). طبق آزمونهای انجام شده، در حالتی‌که درجه ارتباط پرسشها و سند به عنوان تابع هدف در نظر گرفته شود، جواب بدست آمده از BM25 تا حدی بهتر خواهد بود. نتایج به علت نزدیک بودن به BM25 در اینجا ارائه نشده است.

با توجه به مشکل ذکر شده در بالا و موفقیت روش رأی‌گیری وزن‌دار ذکر شده(بوردا)، این روش (بوردا) به عنوان تابع هدف در نظر گرفته شده است. هدف از الگوریتم پیدا کردن وزن‌های متناظر به هر الگوریتم به صورت پویا می‌باشد. در این حالت الگوریتم ComRank نامیده می‌شود. در فعالیت قبلی انجام شده با توجه به اینکه روش نرمال‌سازی متفاوت بود، لذا در نظر گرفتن مقادیر بوردا به صورت مستقیم جواب مناسبی تولید ننمود. بنابراین در آنجا تمام لیست‌ها مرتب شده، و ترتیب جوابهای حاصل روش بوردا به عنوان تابع هدف در نظر گرفته شدند [1]. به عبارت دیگر چونکه مهمترین مسئله در بازیابی اطلاعات ترتیب جواب‌ها هستند نه مقدار آنها، لذا جواب مناسبی در مقایسه با بقیه‌ی روشها حاصل گردید. خواهیم دید روش جدید ارائه شده در اینجا نسبت به حالتی که ترتیب مقادیر را در نظر بگیریم بهتر عمل می‌کند.

در ابتدا با توجه به اینکه الگوریتم ذکر شده مبتنی بر OWA است در ابتدا بردار W محاسبه می‌شود. ویژگیهای در نظر گرفته شده جهت ترکیب شامل الگوریتم‌های رتبه بندی HostRank، BM25، ComLink، Sitmap و TF-IDF می‌باشند. بردار W بدست آمده با استفاده معادله ۱۳ (الگوریتم یادگیری وزن‌های OWA) برابر است با:

$$W = \{0.0013, 0.62, 0.082, 0.29, 0.002\}$$



شکل (۶): مقایسه ComRank با الگوریتم‌های دیگر در P@n

جدول (۲): نمونه‌ای از داده‌های LETOR شامل ویژگیها

Query	PageRank	TF-IDF	BM25	...	Relevancy
qid: 1	0.0023	9.23	19.31	...	1
qid: 1	0.0098	11.3	29.12	...	0
...	...	...	...	...	...
qid: 10	0.000	7.2	15.22		0
...	...	...	...	...	...
qid: 190	0.0076	12.23	30.1		1

برای ارزیابی و مقایسه دو معیار مناسب P@n و NDCG [26] استفاده شده است. معیار P@n نشانگر نسبت تعداد اسناد مرتبط در n سند ارائه شده به n است (معادله (۱۴)).

$$P@n = \frac{\#\text{ of relevant docs in top n results}}{n} \quad (۱۴)$$

با توجه به اینکه در P@n فقط مرتبط بودن یا نبودن یک سند در نظر گرفته می‌شود بنابراین از دقت کافی برخوردار نیست. بنابراین از معیار NDCG (معادله (۱۵)) به عنوان یک معیار تکمیلی نیز استفاده شده است. در این معادله که برای n نتیجه اول استفاده شده است r_j نشان دهنده درجه‌ی مرتبط بودن سند j با پرسش مربوطه می‌باشد.

$$NDCG(n) = \sum_{j=1}^{n} \frac{2^{r_j} - 1}{\log(1+j)} \quad (۱۵)$$

پارامتر دیگری به نام میانگین دقت (AP[39]) نیز در نظر گرفته شده و برای هر پرسش محاسبه می شود. مقدار آن با میانگین P@n برای تمام اسناد مرتبط با پرسش موردنظر(معادله ۱۶) برابر می‌باشد. در این معادله اگر n-امین سند مرتبط باشد، rel(n) برابر ۱ و در غیر این صورت صفر خواهد بود.

$$AP = \frac{\sum_{n=1}^{N} (P@n * rel(n))}{\#\text{ total relevant docs for this query}} \quad (۱۶)$$

همچنین پارامتری دیگری که برای ارزیابی استفاده می‌شود متوسط میانگین دقت (MAP[40]) نام دارد و حاصل میانگین APهای تمام پرسشها می‌باشد.

با توجه به اینکه هر کدام از ویژگیها دارای حوزه مقادیر متفاوتی می‌باشند (جدول (۲))، برای یکدست شدن مقادیر ویژگیها به بازه صفر و یک عمل نرمال سازی به صورت زیر انجام شده است. در این رابطه $f_k(q_i, d_j)$ نشان دهنده ویژگی k برای پرسش qi و سند dj می‌باشد. همچنین $\min\{f_k(q_i, d_l)\}$ و $\max\{f_k(q_i, d_l)\}$ به ترتیب نشان دهنده مقادیر کمینه و بیشنه ویژگی k و پرسش i در تمام اسناد می‌باشند.

$$f_k(q_i, d_j) = \frac{f_k(q_i, d_j) - \min\{f_k(q_i, d_l)\}}{\max\{f_k(q_i, d_l)\} - \min\{f_k(q_i, d_l)\}} \quad (۱۷)$$

قبل ازاینکه به چگونگی یادگیری وزن‌های OWA بپردازیم، ابتدا روش رأی‌گیری ساده به نام بوردا[41] را (ارائه شده در بخش قبل) مورد ارزیابی قرار می‌دهیم. در این روش به هر ویژگی یک وزن ثابت اختصاص داده می‌شود و جمع وزن دار ویژگیها به عنوان خروجی در نظر گرفته خواهد شد. به عبارت دیگر نتایج بر حسب جمع‌وزن‌دار ویژگیها به صورت نزولی مرتب خواهند گردید.

جدول (۳): وزنهای متناظر هر الگوریتم جهت ترکیب که معادل دقت آنها در نظر گرفته شده است.

نام ویژگی	وزن (دقت)	توضیح
PageRank [12]	۰/۰۶	این الگوریتم مبتنی بر گراف وب می‌باشد و در گوگل استفاده شده است.
HostRank [27]	۰/۱۲	این الگوریتم رتبه بندی را باستفاده از ساختار سلسله مراتبی وب انجام می‌دهد.
TF-IDF [8]	۰/۱۴	این الگوریتم مبتنی بر متن Anchor می‌باشد (بهتر از متن جواب می‌دهد).
BM25 [9]	۰/۲۵	الگوریتم احتمالی مبتنی بر کل اسناد می‌باشد.
Sitemap[28]	۰/۲۵	الگوریتم ترکیبی اتصال و محتوا که بر روی سایتها عمل می‌کند.
ComLink [29]	۰/۲۸	الگوریتم ترکیبی اتصال و محتوا بر روی کل گراف وب می‌باشد.

حال مسئله‌ای که وجود دارد پیدا کردن وزن هر ویژگی جهت ترکیب می‌باشد. بدین منظور، مقدار P@n هر ویژگی در داده‌های آموزش (که از قبل محاسبه شده است) به عنوان وزن آن در نظر گرفته شده است. به عبارت دیگر با توجه به اینکه P@n درجه مناسب بودن هر ویژگی (الگوریتم) را نشان می‌دهد به نظر می‌رسد به عنوان وزن الگوریتم مورد نظر مناسب باشد. الگوریتمهای استفاده شده به عنوان ویژگی به همراه وزن نرمال شده آنها (در بازه‌ی صفر و یک) در جدول (۳) نشان داده شده است. در این جدول دقت نرمال شده هر الگوریتم (فرمول (۱۸)) به عنوان وزن در نظر گرفته شده است.

$$P@n\_weight_i = \frac{P@n_i}{\sum_{j=1}^{6} P@n_j} \quad (۱۸)$$

معادله ۱۹ ترکیب ویژگیهای ذکر شده را نشان می‌دهد که در آن اعداد فوق به عنوان وزن رأی‌گیری در نظر گرفته شده‌اند. ویژگی‌های در نظر گرفته شده متفاوت با ویژگیهای ارائه شده در مقاله [1] می‌باشند.

$$W = .06*F_{PageRank} + .12*F_{HostRank} + 0.14*F_{TF-IDF} + 0.25*F_{BM25} + 0.25*F_{Sitemap} + 0.28*F_{ComLink} \quad (۱۹)$$

لازم به ذکر است که می‌توان ویژگیها را بدون نرمال سازی (Borda) و همچنین با نرمال سازی (BordaNorm) ترکیب کرد که هر کدام خصوصیات خاص خود را دارد. در حالت بدون نرمال سازی خصوصیات آرگومانهای هر الگوریتم مانند فاصله ذاتی آنها حفظ خواهد شد. در عین حال با توجه با اینکه محدوده‌ی الگوریتم‌های مبتنی بر اتصال (بین ۰ و ۱) و محتوا (بین صفر تا ۳۰) متفاوت است، لذا ترکیب آنها بدون نرمال سازی مناسب نخواهد بود.

ترکیب می‌باشد. با توجه به اینکه از اهداف مهم روشهای رتبه بندی داشتن دقت بالا است، ما برای هر الگوریتم دقت متناظرش را که توسط قضاوتهای قبلی بدست امده را به عنوان وزن آن در نظر گرفته‌ایم. معادله ۱۱ فرمول مربوطه را نشان می‌دهد که Fi و Pi به ترتیب نشان دهنده مقدار و دقت ویژگیِ i می‌باشند. در قسمت آزمون نشان داده خواهد شد که ترکیب فوق نتیجه چشمگیری داشته است و حتی از بهترین روشهای ارائه شده موجود نیز به مراتب بهتر عمل می‌کند.

$$W = P_1 * F_1 + P_2 * F_2 + ... + P_n * F_n \qquad (۱۱)$$

بزرگترین مشکل روش بوردا ثابت بودن وزن‌ها می‌باشد. با توجه به اینکه محیط بازایابی اطلاعات کاملاً پویا و به محیط (هم بافت) وابسته است لذا لازم است وزن‌ها به صورت وفقی محاسبه شوند.

یکی از روشهای ترکیب موفق پویا عملگر تجمیع OWA می باشد. بنابراین از عمگر تجمیع OWA برای ترکیب استفاده خواهد شد. برای بدست آوردن وزن‌های بردار OWA از مکانیزم یاد گیری مناسب [24] بر مبنای قضاوت کاربر به صورت زیر استفاده شده است.

فرض کنید m ردیف اطلاعات در دسترس باشد که هر ردیف آن شامل یک بردار با n آرگومان به صورت $A_k = (a_{k1}, a_{k2}, ..., a_{kn})$ و بردارِ متناظر مرتب شده‌ی حاصل A، $B_k = (b_{k1}, b_{k2}, ..., b_{kn})$ که b_{kj} متناظر با j-امین عنصر بزرگ در بردار A است (jامین عنصر بعد از مرتب کردن A). در اینجا آرگومانهای هر ردیف متناظر با مقادیر الگوریتم‌های رتبه‌بندی مختلف به ازای یک پرسش و یک سند می‌باشند. بنابراین n الگوریتم رتبه‌بندی به عنوان ویژگی با m جفت پرسش و سند موجود خواهد بود. در بخش نتایج انواع الگوریتم‌های استفاده شده و چگونگی چینش در یک مجموعه آزمون توضیح داده شده است.

فرض کنید یک فرد خبره به هر ردیف اطلاعاتی (جفت سند و پرسش) یک مقداری به نام d_k اختصاص داده باشد (در اینجا نشانگر درجه مرتبط بودن هر پرسش با سند می‌باشد). این عدد در داده‌های استاندارد TREC [18] که در اینجا استفاده شده است شامل دو مقدار صفر و یک می باشد که صفر نشانگر نامرتبط بودن سند و پرسش و یک نشانگر مرتبط بودن آنها است. این مقدار توسط کاربران خبره مقدار دهی شده است.

هدف ما پیدا کردن یک عملگر OWA یا بردار وزندار OWA به صورت $W = (w_1, w_2, ..., w_n)$ است که فرآیند تجمیع روی داده‌ها را مطابق نظر فرد خبره مدل کند. به عبارت دیگر یک تابع تجمیع با شرط زیر به ازای هر k مورد نیاز می‌باشد که هر ردیف اطلاعاتی مانند بردار Ak را به یک مقدار dk نگاشت کند.

$$f(a_{k1}, a_{k2}, ..., a_{kn}) = \sum_{i=1}^{n} w_i * b_{ki} = d_k \qquad (۱۲)$$

الگوریتم یادگیری وزن‌ها به صورت سناریوی تکراری در شکل (۳) نشان داده شده است. در [24] اثبات شده است که این الگوریتم با پیدا کردن مقادیر وزنی مناسب همگرا خواهد شد. مقدار $0 \le \beta \le 1$ نشانگر نرخ یادگیری است و λ (لاندا) برای محاسبه w_i مطابق معادله (۱۳) به کار می‌رود. همچنین $\hat{d}_k$ حاصل تجمیع داده‌های ردیف k با توجه به وزن‌های فعلی بدست آمده است.

۱) مساوی قرار دادن مقادیر اولیه تمام وزن‌ها ($w_i = \frac{1}{n}$) و همچنین $t=0, \lambda_i(0) = 0, \beta = 0.3, \varepsilon = 0.001$

۲) محاسبه لاندا در زمان $t+1$:

$$\lambda_i(t+1) = \lambda_i(t) - \beta w_i (b_{ki} - \hat{d}_k)(\hat{d}_k - d_k)$$

۳) محاسبه w_i با استفاده از لاندای بدست آمده:

$$w_i = \frac{e^{\lambda_i(t)}}{\sum_{j=1}^{n} e^{\lambda_j(t)}} \qquad (۱۳)$$

۴) اگر $(\hat{d}_k - d_k) > \varepsilon$ باشد به مرحله (۲) می‌رویم.

شکل (۳): الگوریتم یادگیری وزن‌های OWA

بنابراین پارامتر λ_i تعیین کننده وزن‌های OWA است که که به وسیله پسخورد خطای[۳۸] $(\hat{d}_k - d_k)$ بروزآوری می‌شود. الگوریتم فوق آنقدر تکرار می‌شود تا با خطای ε متوقف گردد. آزمایشات نشان داده شده است که نرخ یادگیری $\beta = 0.3$ مناسب می باشد.

همانطور که در شکل نشان داده شده است مقدار اولیه وزن‌ها برابر می‌باشند (1/n) که می‌توان به صورت تصادفی و یا دستی (با توجه به داش قبلی) نیز مقداردهی کرد که به کارهای آینده واگذار می‌گردد.

با توجه به مطالب ذکر شده می توان خصوصیات الگوریتم فوق را در موارد زیر خلاصه کرد:

- مبتنی بر هر دو خصوصیت محتوایی و اتصالی
- مقیاس پذیری: به راحتی می‌توان الگوریتم جدید جهت ترکیب به آن اضافه کرد.
- وفقی بودن: ترکیب را بر اساس پرسش، اسناد و نظر کاربر انجام می‌شود. بنابراین بسته به محیط وزن‌ها نیز عوض می‌شوند.

۵- نتایج بدست‌آمده

برای ارزیابی و تست روش ارائه شده از بسته نرم‌افزاری LETOR [25] شامل مجموعه داده های محک برای یادگیری و ترکیب روشهای مختلف جهت رتبه بندی استفاده شده است. این مجموعه شامل ۴۴ ویژگی استخراج شده از Web TREC 2004,2003 مانند PageRank، HITS، TF-IDF، BM25 و غیره برای ۷۵ پرسش مختلف می‌باشد (حدود ۶۵۰۰۰ ردیف داده برای جفت پرسش و سند). به علاوه نظر خبره (درجه‌ی مرتبط بودن) برای هر پرسش و سند نیز وجود دارد. نمونه‌ای از این داده در جدول (۲) نشان داده شده است.

آزمایش انجام شده شامل دو مرحله‌ی آموزش (یادگیری) و آزمون می‌باشد. در مرحله آموزش از ۵۰ پرسش (شامل ۴۰ هزار جفت پرسش و سند)[۱] استفاده شده و مابقی (۲۵ پرسش) برای آزمون مورد استفاده قرار گرفته‌اند. بنابراین تمام معیارهای ارزیابی تعریف شده در زیر روی داده‌های آزمون انجام خواهد شد.

[۱] برای هر پرسش حدود ۸۰۰ سند وجود دارد.

به‌ازای $\alpha = 1$ به بردار وزن $W = [1 \quad 0 \quad \ldots \quad 0]^T$ می‌رسیم که دارای $orness = 1$ است. به‌همین ترتیب به ازای $\alpha = 0$ به بردار وزن $W = [0 \quad 0 \quad \ldots \quad 1]^T$ می‌رسیم که دارای $Orness = 0$ است. میزان Orness این عملگر میانگین‌گیری مرتب وزن‌دار به‌ازای مقادیر مختلف α بصورت زیر می‌باشد:

$$Orness(W) = \frac{1}{n-1}\sum_{i=1}^{n}(n-i)w_i \qquad (۹)$$

یک نوع دیگر عملگر میانگین‌گیری مرتب وزن‌دار تحت عنوان بدبینانه را می‌توان با استفاده از وزن‌های زیر تعریف نمود[17]:

$$\begin{aligned} &w_1 = \alpha^{n-1}, \\ &w_2 = (1-\alpha)\alpha^{n-2}, \\ &w_3 = (1-\alpha)\alpha^{n-3}, \\ &\ldots, \\ &w_{n-1} = \alpha(1-\alpha), \\ &w_n = (1-\alpha) \end{aligned} \qquad (۱۰)$$

این عملگر می‌تواند بصورت معکوس نیز تعریف گردد:

$$\begin{aligned} &w_n = (1-\alpha); \\ &w_j = w_{j+1}\alpha = w_j(1-w_n), \quad j = 2,\ldots,n-1; \\ &w_1 = w_2(1-w_n)/w_n. \end{aligned}$$

می‌توان به فرآیند تهیه این وزن‌ها بصورت دیگری نیز نگاه کرد. فرض کنید W برداری با بعد n از وزن‌ها باشد که آن‌را بصورت زیر به بردار V با بعد $n+1$ تبدیل کنیم:

$$\begin{aligned} &\forall i = 2,\ldots,n : v_{i+1} = w_i, \\ &v_2 = (1-\alpha)w_1 = w_n w_1, \\ &v_1 = \alpha w_1 = (1-w_n)w_1. \end{aligned}$$

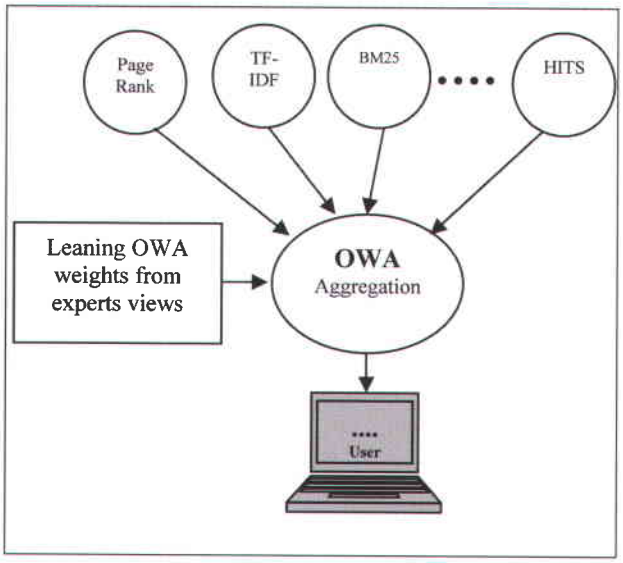
در مرجع [21] نشان داده شده است که این دو نوع عملگر، همه نیازمندی‌های کلی عملگر میانگین‌گیری مرتب وزن‌دار را برآورده می‌سازند و نیز داریم:

$$\forall w_i \in [0,1], i = \{1,\ldots,n\} \quad \sum_{i=1}^{n} w_i = 1$$

۴- الگوریتم ارائه شده (ComRank)

با توجه به مطالب ذکر شده در مقدمه، دقت الگوریتم‌های فعلی کم می‌باشد و بسته به محتوا در شرایط خاص دارای جواب مناسب می‌باشند. بدین منظور سعی شده است با استفاده از ترکیب نتایج حاصل از خصوصیات یک صفحه به عنوان ویژگی[۳۳] با استفاده از عملگر تجمیع OWA الگوریتم ترکیبی جدیدی ارائه شود. با توجه به اینکه هدف، کشف صفحات با کیفیت بالا می‌باشد، بدیهی است که خصوصیات ذاتی صفحه نقش مهمی در کیفیت آن ایفا می‌کند. لذا این خصوصیات به عنوان ویژگیهای یک صفحه در نظر گرفته می شوند. خصوصیات صفحه را می‌توان در دو دسته ریز دانه[۳۴] و درشت دانه[۳۵] دسته بندی کرد. از خصوصیات ریزدانه صفحه می‌توان TF، IDF، تعداد پیوندهای ورودی (In-Degree) و اندازه صفحه (DL) را نام برد. همچنین الگوریتمهای رتبه‌بندی مختلف مبتنی بر اتصال و محتوا مانند PageRank، BM25 و TF-IDF را می‌توان به عنوان ویژگی درشت دانه یک صفحه در نظر گرفت. در دو روش ارائه شده در این مقاله که به ترتیب ComRank و ComRank-FG نام دارند، روشهای درشت دانه و ریز دانه هر یک به صورت جداگانه برای رسیدن به راه‌حل مناسب‌تر ترکیب شده اند. در ادامه ابتدا ویژگیهای درشت‌دانه را با هم ترکیب کرده و بعد از ارزیابی آنها ترکیب ویژگیهای ریز دانه را مورد بررسی قرار می‌دهیم.

شکل (۲) نمای کلی الگوریتم جهت ترکیب چندین ویژگی را نشان می‌دهد.



شکل (۲): نمای کلی ترکیب روشهای رتبه بندی به عنوان ویژگیهای درشت دانه

همانطور که مشاهده می شود، الگوریتم‌های مختلف مبتنی بر معیارهای محتوا و اتصال با استفاده از OWA که وزن‌های آن با استفاده از نظر کاربر خبره بدست آورده شده است، ترکیب می شوند.

الگوریتم‌های رتبه‌بندی موجود علاوه بر داشتن دقت پایین دارای مشکلات دیگری نیز می‌باشند. قرارگرفتن همیشگی صفحات محبوب[۳۶] در صدر لیست ارائه‌شده به کاربر، باعث می‌شود تا کاربر فقط صفحات خاصی را ببیند و در نتیجه صفحات تازه متولد شده با کیفیت بالا که کسی به آنها اشاره نمی‌کند نتوانند در دید کاربران قرار گیرند. این مشکل که "غنی‌تر شدن اغنیاء" نام دارد باعث می‌شود صفحات محبوب مرتباً محبوب‌تر شده و تعداد پیوند به آنها افزایش یابد. برای مثال الگوریتمهای مبتنی بر اتصال مانند PageRank از مشکل "غنی‌تر شدن اغنیاء"[۳۷] [22] رنج می‌برند و به پهنش رتبه بندی حساسیت کمتری دارند. در حالیکه الگوریتمهای مبتنی بر اتصال مانند BM25 و TF-IDF به مشکل فوق حساسیت نداشته ولی دارای مشکل پهنش رتبه می‌باشند. مسلماً ترکیب الگوریتم‌های فوق و ارائه یک الگوریتم مبتنی بر محتوا و اتصال بصورت توأم باعث می‌شود تا مشکلات یکدیگر را تحت پوشش قرار دهند.

مهمترین مسئله باقیمانده چگونگی ترکیب ویژگیهای مختلف می‌باشد. جهت ترکیب از قضاوتهای قبلی کاربر با توجه به پرسش‌ها و اسناد متناظر استفاده شده است.

ساده‌ترین روش، ترکیب وزن دار روشهای رتبه بندی مختلف می باشد که بوردا [23] دارد. در بوردا به هر الگوریتم یک وزن (حق رأی) داده می‌شود. بنابراین مهمترین مسئله در ترکیب، بدست آوردن وزن هر

مساوی ۷ مقداردهی می‌شود. در آزمایشات انجام شده در این مقاله مقادیر k1، b با ۲/۵ و ۰/۸ و k2 و k3 با صفر مقدار دهی شده‌اند[15]. همانطور که دیده می شود الگوریتم فوق از ویژگی‌های ریزدانه مانند فرکانس واژه (tf)، طول سند (dl) و غیره تشکیل شده است.

الگوریتم ComLink

خانم شاکری در [29] یک روش رتبه‌بندی با استفاده از ترکیب پیوند و محتوا ارائه کرده است که در این مقاله ComLink نامیده شده است. در این مقاله یک مدل انتشار وابستگی[۲۶] بین صفحات ارائه شده است. در این مدل امتیازی[۲۷] به نام اَبَر وابستگی[۲۸]، برای هر سند p تعریف شده است که به سه پارامتر شباهت بین پرس‌وجو و سند p ((S(p))، جمع وزن‌دار اَبَر وابستگی صفحاتی که به p اشاره کرده و همچنین p به آنها اشاره می‌کند، بستگی دارد. ترکیب خطی این سه پارامتر به صورت زیر محاسبه می‌شود:

$$h(p)=\alpha S(p)+\beta\sum_{p_i\to p}h(p_i)WI(p_i,p)+\gamma\sum_{p\to p_j}h(p_j)WO(p,p_j) \qquad (۵)$$

$$\alpha+\beta+\gamma=1$$

در معادله فوق، WI و WO به ترتیب وزن پیوندهای ورودی و خروجی صفحه p می‌باشند.

الگوریتم SiteMap

خانم سانگ در مرجع [28] بر خلاف مدل فوق که گراف انتخاب شده مستقل از سایت است، یک روش ترکیبی انتشار مبتنی بر سایت به نام SiteMap ارائه کرده است. همچنین به جای انتشار امتیاز هر صفحه بر حسب میزان شباهت پرس‌وجو و صفحات، فرکانس تکرار واژه‌های هر پرس‌وجو در صفحات انتشار می‌یابد. ابتدا درخت یک وب سایت با استفاده از آنالیز URLها ساخته شده و سپس فرکانس واژه‌های پرس‌وجو از طریق پیوندها از پدر به فرزند (یا فرزند به پدر) در درخت سایت انتشار می‌یابد. معادله زیر روش محاسبه فرکانس تکرار واژه t را در صفحه p نشان می‌دهد.

$$f_t'(p)=(1+\alpha)f_t(p)+\frac{(1-\alpha)}{|Child(p)|}\sum_{q\in Child(p)}f_t(q) \qquad (۶)$$

در این معادله $f_t(p)$ و $f_t'(p)$ به ترتیب تعداد تکرار واژه t در صفحه p را قبل و بعد از انتشار نشان می‌دهند. این روش انتشار بر مبنای سایت[۲۹] نامیده می‌شود. روش فوق بازگشتی نبوده و فقط یکبار اجرا می‌شود (فقط اثر یک همسایه دیده می‌شود).

۳- عملگر تجمیع OWA

بصورت رسمی، تجمیع[۳۰] عبارتست از نگاشت یک n تایی[۳۱] از اشیاء عضو مجموعه‌ای خاص به یک شی از همان مجموعه. در خصوص عملگر ریاضی تجمیع، این مجموعه، همان مجموعه اعداد حقیقی است. در این حالت، این عملگر، تابعی است که یک n تایی از اعداد حقیقی مثل $(x_1,x_2,...,x_n)$ را به یک عدد حقیقی چون y، نسبت می‌دهد:

$$y=Aggreg(x_1,x_2,...,x_n)$$

در سال ۱۹۸۸ توسط آقای یاگر، انواع جدیدی از عملگرهای تجمیع موسوم به عملگرهای میانگین‌گیری مرتب وزن‌دار معرفی شد [17]. عملگر میانگین‌گیری مرتب وزن‌دار از بُعد n، یک نگاشت $F:\Re^n\to\Re$ با یک بردار وابسته n تایی $W=[w_1,w_2,...,w_n]$ با شرط $\left\{\begin{matrix}w_j\in[0,1]\\ \sum_j w_j=1\end{matrix}\right\}$ می‌باشد بطوری‌که:

$$F(a_1,a_2,...,a_n)=\sum_{j=1}^{n}w_jb_j \qquad (۷)$$

که در آن، b_j، برابر با j امین عنصر بزرگ در میان a_iها می‌باشد.

برای بدست آوردن وزن‌ها در عملگر میانگین‌گیری مرتب وزن‌دار، راه‌حل‌های متعددی از جمله توسط O'Hagan مطرح شده است که نیازمند حل یک مساله برنامه‌ریزی غیر خطی مقید[۳۲] است[20]. در کلاس عملگرهای میانگین‌گیری مرتب وزن‌دار نمایی، رابطه بسیار ساده‌ای بین درجه Orness و پارامتر تعیین‌کننده وزن‌های میانگین‌گیری مرتب وزن‌دار وجود دارد. درجه Orness نمایانگر این است که عملگر تجمیع تا چه حد به مقدار بیشینه یا کمینه نزدیک می‌شود. به عبارت دیگر، Orness نزدیک به یک نشانه نزدیکی تابع به عملگر MAX و بالعکس Orness نزدیک به صفر، نمایانگر نزدیکی تابع تجمیع به عملگر MIN است. بطور کلی دو روش برای محاسبه وزن‌ها وجود دارد.

در روش خوش‌بینانه، وزن‌های میانگین‌گیری مرتب وزن‌دار بر اساس فرمول زیر محاسبه می‌شوند:

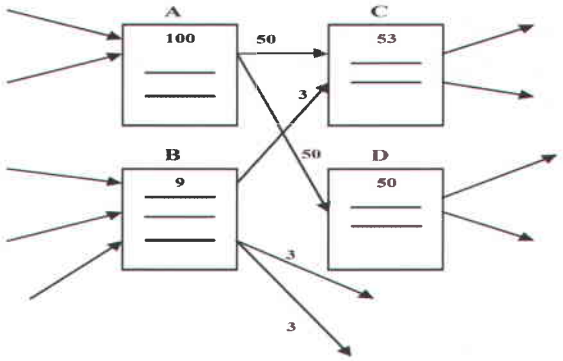
$$\begin{aligned}&w_1=\alpha,\\&w_2=\alpha(1-\alpha),\\&w_3=\alpha(1-\alpha)^2,\\&...\\&w_{n-1}=\alpha(1-\alpha)^{n-2},\\&w_n=(1-\alpha)^{n-1}\end{aligned} \qquad (۸)$$

که در آن پارامتر $\alpha\in[0,1]$ است. می‌توان این وزن‌ها را بصورت بازگشتی زیر نیز تعریف نمود:

$$\begin{aligned}&w_1=\alpha\in[0,1]\\&w_j=w_{j-1}(1-w_1),\quad j=2,...,n-1\\&w_n=w_{n-1}(1-w_1)/w_1\end{aligned}$$

می‌توان به فرآیند تهیه این وزن‌ها بصورت دیگری نیز نگاه کرد. فرض کنید W برداری با بعد n از وزن‌ها باشد که آن‌را بصورت زیر به بردار V با بعد $n+1$ تبدیل کنیم:

$$\begin{aligned}&\forall i=1,2,...,n-1:\ v_i=w_i,\\&v_n=\alpha w_n=w_1w_n,\\&v_{n+1}=(1-\alpha)w_n=(1-w_1)w_n.\end{aligned}$$



شکل (۱): مثالی از PageRank [Page98]

فرمول فوق برای حالتی که گراف کاملاً پیوسته باشد (هر گره به تمام گره‌ها دسترسی داشته باشد) مناسب است. در صورتیکه گراف وب پیوسته نبوده یا صفحات بدون ورودی یا خروجی[18] موجود باشند، الگوریتم مذکور دچار اشکال می‌گردد (الگوریتم همگرا نخواهد شد). به عبارت دیگر بعد از اجرای کامل الگوریتم تعداد زیادی از صفحات دارای مقدار PageRank صفر خواهند بود. برای حل این مشکل از پارامتر d به نام ضریب استهلاک[19] به صورت زیر استفاده شده‌است که در آن n نشان‌دهنده تعداد صفحات وب است.

$$r(i) = (1-d)/n + d * \sum_{j \in B(i)} r(j)/N(j) \qquad (۲)$$

بنابراین هر صفحه به تمام صفحات با احتمال (1-d)/n یک پیوند خواهد داشت. مقدار d در الگوریتم استفاده شده در گوگل ۰/۸۵ مقدار دهی شده است. بدیهی است که رتبه‌بندی حاصل از PageRank رابطه مستقیمی با درجه‌ورودی صفحات دارد. قابل توجه است که درجه‌ورودی، یک ویژگی ریزدانه است.

الگوریتم HostRank

الگوریتم HostRank [27] که مانند PageRank مبتنی بر اتصال می‌باشد هر دو ساختار سلسله مراتبی و اتصالی وب را در رتبه‌بندی لحاظ می‌کند. در این روش صفحات ابتدا در یک ساختار سلسله مراتبی دایرکتوری، هاست[20] و یا دامنه[21] که گره برتر[22] نامیده می‌شود قرار داده می‌شوند و عمل آنالیز اتصال بر روی گراف بدست آمده انجام می‌شود[1]. سپس درجه‌ی اهمیت (ارزش) محاسبه شده‌ی هر گره بین صفحاتی که آن گره شامل آنها می‌شود توسط ساختار سلسله مراتبی توزیع می‌شود. نتایج آزمایشات روی TREC03 و TREC04 افزایش چشمگیری را نسبت به روش‌های دیگر مبتنی بر اتصال مانند PageRank و روش‌های دیگر مبتنی بر ساختار سلسله مراتبی نشان می‌دهند.

الگوریتم TF-IDF

روش وزن‌دهی TF-IDF ارائه شده توسط آقای سالتون [8] از تکرار کلمات سند و پرس‌وجو برای محاسبه وزن استفاده می‌کند. هدف از TF[23] یا تکرار کلمه این است که در صورتیکه یک کلمه چندین‌بار در یک سند ظاهر شود، آن کلمه سند را بهتر توصیف می‌کند. TF معمولاً با توجه به طول سند نرمال می‌شود و فرکانس کلمه بر فرکانس کلمه با بیشترین تکرار تقسیم خواهد شد (tft, d نشان‌دهنده TF کلمه t در سند d می‌باشد). در فرمول (۳) TF-IDF واژه t را در سند d نشان می‌دهد.

$$TF - IDF_{t,d} = tf_{t,d} Idf_t \qquad (۳)$$

$$tf_{t,d} = \frac{freq_{t,d}}{\max_\ell freq_{\ell,d}}$$

$$idf_t = \log \frac{N}{n_t}$$

متغیر IDF[24] نشان‌دهنده عکس تکرار یک کلمه در کل اسناد می‌باشد. پارامتر N نشان‌دهنده تعداد اسناد و nt تعداد اسنادی که شامل کلمه t هستند را نشان می‌دهد.

الگوریتم BM25

آقای رابرتسون [9] یک مکانیزم وزن گذاری مبتنی بر "۲-پواسون" به نام آکاپی[25] که دارای ویرایشهای متنوعی به صورت BMnn است را ارائه کرده است. نامگذاری این مکانیزم (۲-پواسون) به خاطر این است که توزیع هر واژه در یک مجموعه اسناد دارای توزیع پواسون می‌باشد (بدیهی است که تکرار هر واژه در هر مجموعه‌ی اسناد یک توزیع دوجمله‌ای است و در شرایطی که n زیاد و P کم باشد، توزیع دوجمله‌ای به پواسون تبدیل می‌شود). این روش احتمالی از بهترین روشهای رتبه‌بندی به شمار می‌رود که طبق آزمایشات انجام شده دارای دقت حدود ۲۵٪ می‌باشد[16].

جدول (۱): متغیرهای استفاده شده در BM25

نام متغیر	تعریف
R	تعداد اسناد مرتبط به پرسش Q شامل واژه t
R	تعداد اسناد مرتبط به پرسش Q
n	تعداد اسناد شامل واژه t
N	تعداد کل اسناد
tf	فرکانس واژه t در سند d_i
qtf	فرکانس واژه t در سند پرسش
$avdl$	میانگین طول اسناد
dl	طول سند (تعداد واژه‌های در سند)
nq یا $\|Q\|$	تعداد واژه‌های در پرسش
$b.k3 \quad .k2.k1 \quad K$ $K = k_1((1-b) + b(dl_i / dl_{avg}))$	ثابت‌های قابل تنظیم (جهت tuning)
S(Q,D)	شباهت میان سند D و پرسش Q

در الگوریتم BM25 مقدار عددی حاصل برابر جمع وزن همه واژه‌های در پرسش در سند متناظر است و به شکل زیر محاسبه می‌شود (همه متغیرها در جدول (۱) توضیح داده شده‌اند).

$$S(Q,D) = \sum_{i \in Q} (\log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)}) \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} + \qquad (۴)$$

$$k_2 |Q| \frac{avdl-dl}{avdl+dl}$$

لازم است پارامترهای ثابت در نظر گرفته شده در فرمول فوق برای رسیدن به دقت بالا به صورت مناسب مقداردهی شوند. آزمایشات نشان می‌دهد که بسته به مجموعه‌ی داده استفاده شده بعضی از پارامترهای فوق متفاوت خواهند بود. طبق گفته‌ی آقای رابرتسون [19] k1 و b به صورت پیش‌فرض ۱/۲ و ۰/۷۵ مقدار دهی می شوند. k2 و k3 اغلب صفر مقدار دهی می‌شوند ولیکن برای پرسشهای با طول بزرگ، k3 اغلب

[1] برای مثال yazduni.ac.ir، pweb.yazduni.ac.ir و pweb.yazduni.ac.ir/staff به ترتیب دامنه، هاست و دایرکتوری می‌باشند.

مهم از این مجموعه عظیم داده‌ای، موتورهای جستجو می‌باشند. لذا حدود ۸۰٪ از کاربران اینترنت سایتهای جدید را از طریق موتورهای جستجو کشف و بازدید می‌کنند[5]. به علاوه تعداد کاربرانی که از موتورهای جستجو استفاده می‌کنند نیز به طور دایم در حال افزایش است. برای مثال میانگین روزانه تعداد کاربران موتور جستجو در آمریکا از ۴۹/۳ میلیون در سپتامبر ۲۰۰۴ به ۶۰/۷ میلیون در سپتامبر ۲۰۰۵ (۲۳٪ رشد) افزایش یافته است [6].

بخش رتبه‌بندی[7] یکی از مهمترین قسمت‌های موتورجستجو می‌باشد. "رتبه‌بندی" فرآیندی است که طی آن کیفیت صفحات از جنبه ارتباط با پرسش کاربر توسط موتور جستجو تخمین زده می‌شود. با توجه به اینکه به ازای هر پرسش کاربر معمولاً هزاران صفحه مرتبط وجود دارد لازم است آنها را اولویت بندی کرده و ۱۰ یا ۲۰ تای اول را به کاربر نشان دهد.

در حال حاضر دو روش عمده رتبه‌بندی مبتنی بر محتوا[8] (استفاده شده در بازیابی اطلاعات سنتی) و مبتنی بر اتصال [9] (استفاده شده در وب فعلی) وجود دارد.

در روشهای مبتنی بر محتوا که در بازیابی اطلاعات سنتی [7] استفاده می‌شوند، مدلهایی مانند بولی[10]، احتمالی و فضای برداری[11] جهت رتبه‌بندی اسناد بر مبنای محتوای آنها ارائه شده‌است. مهمترین این روشها در مدل برداری TF-IDF [8] و در مدل احتمالی BM25 [9] می‌باشند. به علت حجم زیاد محتوای وب مهمترین مشکل الگوریتم‌های بر مبنای محتوا پهنش رتبه[12] [10] می‌باشد.

در حالیکه در حوزه بازیابی اطلاعات وب[13] علاوه بر معیارهای فوق از ساختار اتصالی وب برای رتبه‌بندی استفاده می‌شود. پیوندها[14] بیان‌کننده کیفیت محتوای یک صفحه از منظر صفحات بیرونی می‌باشد (بر خلاف محتوای متنی صفحه که کاملاً به ایجاد کننده آن وابسته است). به عبارت دیگر در رتبه‌بندی بر مبنای پیوند از محتوای صفحات دیگر برای ارزیابی یک صفحه استفاده می‌شود[11]. این خاصیت باعث می‌شود الگوریتم رتبه‌بندی با استفاده از اطلاعات استخراجی از پیوندها به مسائلی مانند پهنش حساسیت کمتری نشان دهد. از الگوریتم‌های مبتنی بر اتصال می‌توان PageRank [12] (اولین الگوریتم استفاده شده در گوگل)، HITS [13] و DistanceRank [14] را نام برد.

در مراجع [15, 16] نشان داده شده است که دو دسته الگوریتم‌های فوق دارای دقت[15] پایین می‌باشند. همچنین نشان داده شده است که الگوریتم‌های مبتنی بر محتوا مانند BM25، الگوریتم‌های مبتنی بر اتصال و روشهای ترکیبی به ترتیب دارای دقت حدود ۲۵٪، ۱۴٪ و ۲۸٪ هستند.

با توجه با اینکه الگوریتم‌های ارائه شده در مجموع دارای کارایی پایینی می باشند و همچنین معمولاً هر الگوریتم در دامنه و محیط خاص خوب عمل می‌کند به نظر می‌رسد با استفاده از ترکیب الگوریتم‌های موجود بتوان علاوه بر حل مشکلات فعلی به کارآیی بهتری نیز دست پیدا کرد.

در این مقاله هدف ارائه یک الگوریتم وفقیِ ترکیبی برای دستیابی به دقت و کارایی بالاتر می‌باشد. این الگوریتم با استفاده از ترکیب الگوریتم‌های موجود به عنوان ویژگیهای درشت‌دانه، همچنین ترکیب ویژگیهای ریزدانه مانند TF، IDF و In-Degree به کمک فرآیند یادگیری سعی خواهد کرد به الگوریتم بهتری دست پیدا کند.

فرآیند یادگیری جهت ترکیب الگوریتم‌های مختلف با استفاده از OWA[16] [17] انجام می‌شود. برای ارزیابی و مقایسه با سایر روشها از مجموعه داده محک TREC 2004 [18] استفاده شده است. نتایج آزمایشات، افزایش قابل توجهی را در مقایسه با بقیه الگوریتم‌ها نشان می‌دهد. الگوریتم ارائه شده‌ی بر مبنای ویژگی‌های درشت‌دانه دارای میانگین دقت حدود ۳۰٪ می باشد که ۶٪ بهتر از بهترین الگوریتم فعلی یعنی BM25 می‌باشد. بزرگترین مشکل این الگوریتم پیچیدگی محاسباتی آن می‌باشد. در حالیکه الگوریتم ترکیبی مبتنی بر ویژگی‌های ریزدانه علاوه بر داشتن دقت مناسب، دارای پیچیدگی خیلی کمی می‌باشد.

این فعالیت در حقیقت ادامه کار انجام شده در [1] می‌باشد. در اینجا علاوه بر ارائه یک الگوریتم ترکیبی مبتنی بر ویژگی‌های درشت‌دانه کامل با جزئیات بیشتر و کارآیی بهتر یک الگوریتم ریزدانه نیز ارائه شده است.

ساختار مقاله بدین صورت می‌باشد که ابتدا در بخش بعدی مهمترین الگوریتمهای پایه موجود بصورت خلاصه تشریح می‌شوند. در بخش (۳) کلیات الگوریتم OWA و خصوصیات آن شرح داده خواهد شد. جزئیات الگوریتم ارائه شده در بخش (۴) ارائه می‌شود و در بخش (۵) نتایج آزمایشات ارائه خواهد شد. در پایان بخش (۶) شامل نتیجه‌گیری و کارهای آینده می‌باشد.

۲- الگوریتم‌های رتبه‌بندی موجود

در این بخش الگوریتم‌های استفاده شده در مقاله را تشریح می‌کنیم.

الگوریتم PageRank

الگوریتم PageRank یک الگوریتم مستقل از پرسش می‌باشد که در موتور جستجوی گوگل استفاده شده‌است و بر اساس اتصال بین صفحات عمل می‌کند. برای مثال اگر صفحه p1 به صفحه p2 اشاره‌کند، موضوع p2 برای ایجاد کننده p1 جذاب می‌باشد. بنابراین تعداد پیوندهای ورودی به یک صفحه درجه جذابیت آن صفحه برای دیگران را نشان می‌دهد. در نتیجه درجه جذابیت یک صفحه با تعداد پیوندهای ورودی آن افزایش می‌یابد. به‌علاوه وقتی به یک صفحه از صفحات مهم (با تعداد پیوند زیاد) اشاره شود، آن صفحه نیز رتبه‌ی بالایی خواهد داشت. به‌عبارت دیگر وزن هر صفحه در PageRank جمع وزن‌دار صفحاتی است که به آن اشاره می‌کنند. بنابراین الگوریتم PageRank بازگشتی بوده و با فرض برقرار بودن برخی شرایط می‌توان آن را با استفاده از زنجیره مارکف مدل کرد.

فرض کنید N(i) و B(i) به ترتیب نشان‌دهنده تعداد پیوندهای خروجی[17] و مجموعه صفحات ورودی صفحه i باشند. رتبه صفحه i با استفاده از PageRank به‌صورت زیر محاسبه می‌شود:

$$r(i) = \sum_{j \in B(i)} r(j) / N(j) \qquad (۱)$$

درنتیجه رتبه صفحه i مساوی جمع رتبه‌های صفحات ورودی تقسیم بر درجه خروجی آنها می‌باشد.

تقسیم ارزش یک صفحه بر درجه خروجی باعث می‌شود تا اولاً رتبه صفحه به صورت عادلانه بین فرزندان (خروجیها) تقسیم شود و ثانیاً جمع وزن‌های همه صفحات خروجی به عدد ثابت (یک) نرمال شود. شکل (۱) مثالی از PageRank را نشان می‌دهد. رتبه صفحات A که ۱۰۰ می باشد بین دو صفحه C و D تقسیم می‌شود. همچنین رتبه صفحه‌ی B نیز بین فرزندانش که از جمله آنها C می‌باشد تقسیم می‌شود. بنابراین صفحه C، ۵۰ رتبه از صفحه A و ۳ رتبه از صفحه B دریافت می‌کند و در مجموعه دارای رتبه‌ی ۵۳ خواهد بود.

مجله بین المللی علمی- پژوهشی
ارتباطات وفناوری اطلاعات

جلد یکم - شماره ۴ - شهریور ماه ۱۳۸۸ (ش ص ۷۰-۵۹)

الگوریتم ترکیبیِ وفقی جهت رتبه‌بندی صفحات وب با استفاده از ویژگیهای ریزدانه و درشت‌دانه

علی محمد زارع بیدکی
دانشگاه یزد
دانشکده مهندسی برق و کامپیوتر
alizareh@yazduni.ac.ir

محمد آزادنیا
مرکز تحقیقات مخابرات ایران
پژوهشکده فناوری اطلاعات
azadnia@itrc.ac.ir

ناصر یزدانی
دانشگاه تهران
دانشکده مهندسی برق و کامپیوتر
yazdani@ut.ac.ir

امیرحسین کیهانی‌پور
دانشگاه تهران
دانشکده مهندسی برق و کامپیوتر
keyhanipour@ut.ac.ir

تاریخ دریافت: ۱۳۸۷/۱۲/۲۰- تاریخ پذیرش: ۱۳۸۸/۶/۲۴

چکیده: حجم عظیم و پویا بودن اطلاعات وب، یکی از مهمترین چالش‌های بازیابی اطلاعات در پاسخ به پرسش کاربر می باشد. برای بهبود نتایج جستجو تاکنون الگوریتم‌های متنوعی مانند BM25 و PageRank ارائه شده‌اند. در این مقاله یک الگوریتم رتبه بندی وفقی[1] ترکیبی برای دستیابی به دقت و کارایی بالاتر ارائه شده است. این الگوریتم با استفاده از ترکیب الگوریتم‌های رتبه‌بندی موجود به عنوان ویژگی درشت‌دانه مانند BM25 و TF-IDF و همچنین ترکیب ویژگیهای ریزدانه‌ی موجود مانند تکرار واژه‌ها[2] و درجه ورودی صفحات[3] به کمک فرآیند یادگیری به کارآیی بهتری دست یافته است. در فرآیند یادگیری از عملگر تجمیع OWA[4] و نظر افراد خبره در مورد درجه ارتباط پرسش و سند استفاده می‌شود. برای ارزیابی الگوریتم پیشنهادی از مجموعه داده‌های محک LETOR[5] شامل داده‌های "WEB TREC 2004"[6] استفاده گردید. آزمایشات، افزایش چشمگیری را در میزان دقت بازیابی نشان می‌دهند.

کلمات کلیدی: الگوریتم‌های رتبه بندی وب، رتبه‌بندی‌ترکیبی، عملگرتجمیع OWA، دقت

۱- مقدمه

توسعه و رشد نمایی وب باعث شده است تا با حجم عظیمی از اطلاعات (میلیاردها صفحه) شامل اسناد با فرمت‌های متفاوت (متن/صوت/تصویرو...) در مکان‌های مختلف مواجه شویم. برای مثال حجم اطلاعات نمایه‌سازی شده در سال ۱۹۹۴ از ۱۱۰۰۰۰ صفحه [2] به ۱۱/۵ میلیارد صفحه در سال ۲۰۰۵ [3] و ۲۶ میلیارد در ۲۰۰۸ رشد یافته‌است[4].

برای اینکه کاربر به راحتی بتواند در این اقیانوس اطلاعاتی جستجو نموده و به اطلاعات مورد نظرش دسترسی پیدا کند، نیاز مبرمی به سامانه‌های بازیابی اطلاعات وجود دارد.

در حال حاضر کاراترین ابزارها برای مدیریت، بازیابی و استخراج اطلاعات