4th International Conference on Web Research
April 25-26, 2018; Tehran, IRAN
ICWR2018
کنفرانس بین المللی وب پژوهی
چهارمین
۵ و ۶ اردیبهشت ۱۳۹۷؛ تهران، ایران

A New Approach to Improve the Accuracy of the TF_IDF Ranking Algorithm in Text Retrieval

[1] Azize Nemati, [2]Soheila Karbasi

[1] Graduate Student, Department of Computer Engineering, Golestan University, Gorgan
azize.nemati94@gmail.com

[2] Assistant Professor, Department of Computer Engineering, Golestan University, Gorgan
s.karbasi@gu.ac.ir

Abstract

Today, the World Wide Web is considered as the largest source of data with the help of Web search engines, as one of the most useful tools for extracting information. Due to the web growth, providing information related to user queries by search engines is very difficult. Also, the effectiveness of the information retrieval systems is largely dependent on term-weighting. Therefore, search engines use different web mining techniques to rank search results. For this purpose, various ranking algorithms are presented.

In this research, the weighting algorithm TF_IDF is used to rank the documents. By introducing the entropy parameter related to the number of user query words in the text of the documents, the accuracy of the ranking of the documents in the information retrieval is evaluated. The remarkable points obtained from the surveys on standard questions provide a new approach to increasing the efficiency of text search systems, which the responses from subsequent experiments demonstrate its validation. The proposed approach in this paper uses the Standard Web collections and the results show that it can significantly increase the accuracy of retrieval in terms of the volume of test data collection.

Keywords: Information retrieval, Web mining, TF_IDF weighting model, Document scoring, Entropy

4th International Conference on Web Research
April 25-26, 2018; Tehran, IRAN
ICWR2018
کنفرانس بین المللی وب پژوهی
چهارمین
۵ و ۶ اردیبهشت ۱۳۹۷؛ تهران، ایران

معرفی رویکردی جدید در بهبود دقت الگوریتم رتبه بندی TF_IDF جهت بازیابی اسناد متنی

عزیزه نعمتی[۱]، سهیلا کرباسی[۲]

[۱] دانشجوی کارشناسی ارشد، گروه مهندسی کامپیوتر، دانشگاه گلستان، گرگان
azize.nemati94@gmail.com

[۲] استادیار، گروه مهندسی کامپیوتر، دانشگاه گلستان، گرگان
s.karbasi@gu.ac.ir

چکیده

امروزه وب گسترده جهانی به عنوان بزرگ‌ترین منبع داده‌ها، به کمک موتورهای جستجوی وب، بعنوان یکی از پر کاربردترین ابزار استخراج اطلاعات به شمار می‌رود. با توجه به رشد روز افزون وب، فراهم کردن اطلاعات مرتبط با پرسجوی کاربر توسط موتورهای جستجو بسیار مشکل شده است. نیاز فعلی موتورهای جستجو آن است که بتوانند اسناد را با بالاترین دقت در اختیار کاربران قرار دهند. بنابراین موتورهای جستجو از تکنیک‌های مختلف وب کاوی برای رتبه بندی نتایج جستجو استفاده می‌کنند. برای این منظور الگوریتم‌های رتبه‌بندی متنوعی ارائه شده است.

در این تحقیق، برای رتبه بندی اسناد از الگوریتم وزنی TF_IDF استفاده می‌شود که با اضافه کردن پارامتر آنتروپی مربوط به تعداد تکرار واژه‌های پرسجوی کاربر در متن اسناد، دقت رتبه بندی اسناد در بازیابی اطلاعات، ارزیابی می‌شود. نکات قابل توجه بدست آمده از بررسی‌های انجام شده بر روی پرسش‌های استاندارد، رویکردی جدید جهت افزایش کارایی سیستم‌های جستجوی متنی را ارائه می نماید که پاسخ های حاصل از آزمایشات بعدی، حاکی از تصدیق آن می‌باشد. به این ترتیب رویکرد پیشنهادی در این مقاله از شاخه محتواکاوی وب استفاده می‌کند و نتایج نشان می‌دهد که می‌تواند به نسبت حجم داده های کلکسیون تست، دقت بازیابی اسناد آن‌را به میزان قابل توجهی افز ایش دهد.

کلمات کلیدی

بازیابی اطلاعات، وب کاوی، مدل وزنی TF_IDF، رتبه بندی اسناد، آنتروپی

۱- مقدمه

امروزه وب گسترده جهانی (www)، یک منبع عظیم از اطلاعات غیر همگن شامل متن، تصویر، صوت، فیلم و متادیتا می‌باشد که بر روی صفحات مختلفی قرار دارند که با یکدیگر پیوند دارند. بر اساس تخمین صورت گرفته، مشخص شده است که وب از زمان پیدایش خود چیزی بالغ بر ۲۰۰۰ درصد گسترش یافته و در طول هر شش ماه اندازه آن دو برابر می‌شود[1]. اکثر کاربران از ابزارهای بازیابی اطلاعات مانند موتورهای جستجو، برای دستیابی به اطلاعات مورد نیاز خود استفاده می‌کنند. وب کاوی[۱] یکی از کاربردهای داده کاوی است که در آن به کشف الگوها در وب پرداخته می‌شود.

وب کاوی با استفاده از تکنیک های داده کاوی به دنبال خودکارسازی، کشف و استخراج اطلاعات مفید از اسناد وب و سرویس های آن می باشد. در این میان هدف اصلی، استنتاج اطلاعات با ارزش و مفید از صفحات وب می باشد. هر موتور جستجو از اجزاء مختلفی تشکیل شده است که در این میان الگوریتم رتبه دهی یک موتور جستجو، از مهم‌ترین اجزای تعیین کننده توانایی و کیفیت یک موتور جستجو می‌باشد[2].

رتبه بندی نتایج جستجو، یکی از مشکلات اساسی در سیستم‌های بازیابی اطلات است. با توجه به پرسش کاربر و در دست داشتن D سند مرتبط، مسئله اصلی مرتب کردن این اسناد است به‌طوری که بهترین نتایج در ابتدای لیست نتایج، به کاربر نشان داده شود. در این مقاله ابتدا در بخش ۲ مفاهیم اولیه و مهم در فرایند بازیابی اطلاعات توسط موتورهای جستجو بیان می‌شود. در بخش ۳، نگاهی به کارهای مرتبط در این زمینه می‌اندازیم. در بخش ۴ رویکرد پیشنهادی بیان شده است. در بخش ۵ به ارزیابی کارایی و نتایج مدل وزنی پیشنهادی پرداخته می‌شود و در نهایت در بخش ۶ به بیان نتایج و پیشنهادهایی برای کارهای آینده می‌پردازیم.

۲- حوزه‌های وب کاوی

وب‌کاوی شامل استخراج اطلاعات در وب، و به دنبال آن تعمیم و تحلیل این اطلاعات است که می توان آن را داده کاوی در وب نامید. تحقیقات انجام شده در وب کاوی را می توان به سه بخش ساختار کاوی وب[۲]، محتواکاوی وب[۳] و کاربردکاوی وب[۴] تقسیم نمود. در ساختارکاوی وب، هدف تحلیل ساختار پیوندهای وب با استفاده از مبانی تئوری گراف است. به این ترتیب وب به صورت گرافی در نظر گرفته می‌شود که در آن هر گره متناظر با یک صفحه وب، و هر یال بین دو گره، به عنوان یک پیوند بین دو صفحه متناظر می‌باشد. در محتواکاوی وب اطلاعات مفید از متن و محتوای وب استخراج می‌شود. محتوا می‌تواند متن، صوت، فیلم و یا هر چیز دیگری باشد. از آن جا که محتوای وب دارای منابع ساخت یافته و نیمه ساخت یافته و داده‌های چند رسانه‌ای می‌باشد، کاوش محتوای وب، کاری سخت و پیچیده است. در کاربرد کاوی وب به استخراج و تحلیل الگوی علایق کاربران در استفاده از وب پرداخته می‌شود. برای این کار به عنوان مثال از تعداد لاگ‌های هر صفحه استفاده می‌شود [4,5].

در هر یک از حوزه‌های معرفی شده در وب، مشکلات پویایی، رشد و تغییرات سریع در داده‌های وب، وجود دارد که این امر ناشی از وجود داده‌های موقتی می‌باشد. بطور یقین، وب به عنوان بزرگ‌ترین منبع داده‌ها می‌تواند توسط موتورهای جستجوی وب، به صورت یکی از پر استفاده‌ترین ابزارهای استخراج اطلاعات در اینترنت، به سهولت مورد دسترسی قرار گیرد. اما رشد تصاعدی و آهنگ سریع تغییر و تحول در وب، بازیابی اطلاعات مرتبط را پیچیده می‌سازد. به دلیل اهمیت موارد ذکر شده، در این پژوهش به بررسی رتبه بندی اسناد مورد بازیابی و راهکارهای مربوط به بهبود عملکرد آنها پرداخته شده است.

۳- انواع الگوریتم های رتبه بندی صفحات وب

انواع الگوریتم هایی که برای رتبه بندی استفاده می شود عبارتند از:

۱. الگوریتم PageRank (PR) [2,6,7]
۲. الگوریتم Weighted PageRank (WPR) [8,9]
۳. الگوریتم Page Content Rank (PCR) [10]
۴. الگوریتم Hyperlink Induced Topic Search (HITS) [6,11,12,13]
۵. الگوریتم Spamming Resistant Expertise (SPEAR) [14]
۶. الگوریتم Visit Of Link based page link (VOL) [15,16]
۷. الگوریتم Sim Rank [17]
۸. الگوریتم مبتنی بر اولویت بندی کاربر (User Preference Based Page Ranking Algorithm) [18]

هر یک از الگوریتم‌های رتبه بندی ممکن است از یک حوزه و یا ترکیب چند حوزه‌ی وب‌کاوی استفاده کنند. در ادامه به معرفی برخی از پرکاربردترین این الگوریتم‌ها می‌پردازیم.

سال ۱۹۹۸ برای الگوریتم‌های رتبه‌بندی موتورهای جستجو بسیار مهم بود و انقلابی در الگوریتم های رتبه بندی ایجاد شد. در همین سال بود که دو الگوریتم HITS و PageRank معرفی شدند [2,7]. در سال ۲۰۰۴ زینگ و قربانی الگوریتم Weighted PageRank را معرفی کردند که توسعه یافته الگوریتم PageRank است. Weighted PageRank اهمیت لینک‌های ورودی و لینک‌های خروجی صفحات را به حساب می آورد و امتیازهای رتبه‌بندی را مبنی بر محبوبیت صفحات توزیع می‌کند. این الگوریتم توانایی شناسایی تعداد بیشتری از صفحات مربوط به یک پرسجوی داده شده را در مقایسه با PageRank استاندارد دارد [8,9].

در سال ۲۰۱۰ یک نوع جدید از الگوریتم رتبه‌بندی بوسیله ترکیب درخت طبقه‌بندی شده با الگوریتم ایستای رتبه بندی صفحه PageRank پیشنهاد شد، که قادر است درخت طبقه بندی شده را مطابق با تعداد زیادی از نتایج جستجوی مشابه کاربر ایجاد کند[19].

در الگوریتم Agent based Weighted PageRank برای ساختار کاوی وب همچنین محتوا کاوی وب از Agentها استفاده می‌شود. این الگوریتم تحلیل‌هایی را بر روی درخواست کاربران انجام می‌دهد و مشخص می‌کند که کاربر به دنبال چه چیزی است تا صفحات مرتبط و مهم در لیست نتایج ظاهر شوند [20].

الگوریتم WPRCLV نسخه توسعه یافته‌ای از WPR است که از محتوای صفحات وب و بازدید لینک‌ها استفاده می‌کند[21]. در این الگوریتم معماری موتور جستجو با اضافه شدن پارامتر ارتباط صفحه و رفتار کاربران در حین جستجوی وب و همچنین افزونه‌ای برای محاسبه اهمیت صفحات، اصلاح می‌شود.

در سال ۲۰۱۸ الگوریتم جدیدی مبتنی بر چارچوب محاسبات ابری که بر پایه Hadoop-map Reduction می‌باشد ارائه شده است که می‌تواند به صورت Meta search و ابزار رتبه بندی صفحات اجرا شود[3].

الگوریتم SimRank یکی از الگوریتم‌های نسبتا جدید است که بر اساس معیار شباهت[۵] و همانندی در محیط فضای برداری عمل می‌کند و می‌تواند برای رتبه بندی نتایج پرسجو از صفحات وب، به صورت موثر و کارایی مورد استفاده قرار گیرد. به طور کلی الگوریتم‌های رتبه بندی سنتی، فقط به رابطه میان لینک‌های بین صفحات وب برای محاسبه رتبه توجه داشتند، اما دقت مربوط به رتبه صفحات تا میزان زیادی به محتوای صفحات وابسته است.

اساس کار الگوریتم SimRank بر پایه‌ی وزن‌دهی واژه[۶] است[17] که مدل وزنی TF_IDF یکی از پرکاربردترین مدل‌ها است که در این الگوریتم استفاده شده است.

۳-۱- مدل وزنی TF_IDF

TF_IDF مخفف (فرکانس ترم- معکوس فرکانس سند) بوده و امروزه یکی از محبوب‌ترین مدل‌های وزن‌دهی در سیستم‌های بازیابی اطلاعات است و ۸۳٪ از سیستم‌های پیشنهاد دهنده[۷] در حوزه کتابخانه‌های دیجیتال از آن استفاده می‌کنند[23].

در الگوریتم‌های رتبه‌بندی مبتنی بر این مدل برای هر واژه از مقادیر کمی دو پارامتر TF(Term Frequency) که بیانگر دفعات تکرار واژه در سند و IDF(Inverse Document Frequency) که نشان دهنده معکوس تعداد اسنادی است که آن واژه در آنها ظاهر می‌شود استفاده شده است.

مقدار TF واژه t_i در سند d_j با استفاده از رابطه (۱) نرمال سازی می‌شود:

$$TF_{ij} = \frac{fij}{\max\{f1j, f2j, \ldots\ldots, f|v|j\}} \qquad (۱)$$

که در آن f_{ij} نشان دهنده تعداد تکرار واژه t_i در سند j می‌باشد. $|v|$ برابر تعداد واژگان موجود در صفحه است [17].

مقدار IDF مربوط به واژه t_i با استفاده از رابطه (۲) محاسبه می‌شود که N تعداد کل اسناد موجود در مجموعه است و df_i تعداد صفحاتی است که واژه t_i حداقل یک بار در آنها ظاهر می‌شود.

$$IDF_i = \log\frac{N}{df_i} \qquad (۲)$$

در نهایت وزن واژه توسط رابطه (۳) حاصل می‌شود:

$$W_{ij} = TF_{ij} \cdot IDF_i \qquad (۳)$$

موتورهای جستجوی فعلی عموما نتایج را در تعداد زیادی از صفحات در اختیار کاربر قرار می‌دهند، در حالی که کاربر انتظار دارد در کمترین زمان ممکن به بهترین نتایج برسد، بدون آنکه مجبور شود برای یافتن نیاز خود در میان مجموعه‌ای از تمام صفحات به جستجو بپردازد. با توجه به اینکه امروزه بیشتر موتورهای جستجو از انواع مختلفی از معیارهای TF_IDF برای رتبه بندی اسناد استفاده می‌کنند، در راهکار پیشنهادی این مقاله با بررسی پرسش-های موجود در کلکسیون‌های استاندارد و محاسبه آنتروپی تکرار واژه‌های پرسش کاربر در اسناد، پارامتری را معرفی می‌کنیم که استفاده از آن در مدل وزنی مورد استفاده، دقت بالاتری برای نتایج جستجو در الگوریتم رتبه‌بندی را فراهم می‌نماید.

٤-رویکرد پیشنهادی

در این پژوهش مجموعه اسنادی که برای هر پرسش به عنوان برترین اسناد مرتبط برای هر پرسش کلکسیون معرفی شده‌اند، مورد توجه قرار می‌گیرد. ما این اسناد را در نرم افزار متلب از نظر تعداد تکرار کلمات هر پرسش مورد تجزیه قرارداده و در پی آن هستیم که پارامترهای موثری را که در این اسناد وجود دارد و باعث شده است تا این اسناد به عنوان سندهایی با رتبه برتر برای آن پرسش انتخاب شوند را مورد بررسی قرار دهیم.

در این جا اسناد مربوطه و پرسش‌ها مورد پیش‌پردازش قرار می گیرند و ایست‌واژه‌ها (واژه‌هایی که در بیشتر اسناد هستند و تاثیری در وزن‌دهی اسناد ندارند) حذف می‌شوند. سپس الگوریتم Porterstemmer (الگوریتم ریشه یابی واژه‌ها) را به صورت تابع به همه کلمات اعمال می کنیم و در نهایت تعداد تکرار کلمات هر پرسش در اسنادی که رتبه آن ها به صورت نزولی مرتب شده اند، مشخص می‌شود.

٤-۱- استخراج پارامتر موثر بر رتبه‌بندی اسناد

ایده اصلی پژوهش حاضر این است که نحوه توزیع کلمات پرسش می‌تواند در انتخاب اسناد موثر باشد. به این معنی که هر چه توزیع کلمات پرسش در سندی منظم‌تر باشد، آن سند باید رتبه بالاتری را کسب کند. بنابراین با فرض این ایده به بررسی پارامترهای مختلف می‌پردازیم.

برای ارزیابی تاثیرگذاری پارامتر آنتروپی معرفی شده، ابتدا این پارامتر برای پرسش‌های هر مجموعه تست، بر روی مجموعه اسناد موجود در فایل داوری[۸] هر سند که بر اساس معیار بهترین سندها مرتب شده‌اند، محاسبه می‌شود. برای نشان دادن روش کار به صورت نمونه در جدول (۱) مراحل محاسبه این پارامتر و مقدار آن برای پرسش ۱۸ از مجموعه CACM نشان داده شده است.

در جدول (۱) سطرها شامل ریشه واژه‌های موجود در پرسش می‌باشند و در ستون‌ها هم اسناد به ترتیب رتبه آنها در فایل داوری قرار گرفته‌اند. ملاحظه می‌شود که وقتی تعداد تکرار واژه پرسجو در دو یا چند سند برابر می-شود، سندی بهتر است که آنتروپی کمتری دارد. در اینجا با اینکه سند ۷ از نظر فرکانس واژه‌ها تعداد بیشتری نسبت به سند ۱ برخوردار است، اما رتبه کمتری را در داوری کسب نموده است. علت این عامل هم آن است که سند ۱ آنتروپی کمتری نسبت به سند ۷ دارد. بررسی‌های بیشتر نشان از تاثیر گذاری این فاکتور در بیشتر اسناد دارد. البته در این میان فاکتورهای دیگری

جدول (۱):محاسبه پارامتر آنتروپی پرسش ۱۸ از مجموعه CACM

Term of Query	Doc1 1158	Doc2 1215	Doc3 1262	Doc4 1471	Doc5 1613	Doc6 1811	Doc7 2060
code	0	0	0	0	1	0	0
compact	0	0	0	0	0	0	0
compil	0	4	0	0	4	0	0
especi	0	0	0	0	0	0	0
highli	0	0	0	0	0	0	0
horizont	0	0	0	0	0	0	0
languag	0	0	2	3	0	0	7
machin	0	0	0	0	0	0	0
microcod	0	0	0	0	0	0	0
parallel	2	0	4	1	3	7	2
processor	0	0	0	0	4	16	0
entropy	0	0	0.384	0.339	0.776	0.370	0.319
Num_word	25	29	45	48	51	94	102

چون طول سند نیز تاثیرگذار می‌باشد.

در جدول (۲) این پارامتر برای پرسش ۱۰ از کلکسیون Medline محاسبه شده است.

لازم به ذکر است که همان‌گونه که در جدول (۲) مشاهده می‌شود، برای واژه-های با تکرار مساوی غیر صفر، سندی دارای رتبه بالاتر است که آنتروپی کمتری داشته باشد. البته این پارامتر هم شامل موارد استثنا می‌شود، چرا که در این میان پارامترهای دیگری چون طول سند نیز موثر واقع می‌شود.

با دقت در نتایج کلی حاصل از همه پرسش‌ها در هر سه مجموعه، مشخص می‌شود که در ۸۰٪ موارد، در میان اسنادی که تعداد تکرار برابری برای یک واژه پرس و جو دارا می‌باشند، سندی بهتر بوده که آنتروپی کمتری دارد. بنابراین می‌توان نتیجه گرفت در این موارد، رتبه یک سند با پارامتر آنتروپی که قبلا معرفی شد، نسبت عکس دارد.

جدول (۲):محاسبه پارامتر آنتروپی پرسش ۱۰ از مجموعه Medline

Term of Query	Doc1 93	Doc2 94	Doc3 96	Doc4 141	Doc5 173	Doc6 174	Doc7 175
bronchial	0	0	0	0	0	0	0
culture	5	4	0	0	1	1	5
lung	2	1	0	3	1	0	0
neoplasm	0	0	4	0	0	0	0
tissu	0	1	0	0	1	0	4
Entropy	0.536	0.777	0	0	0.985	0	0.615
Num_word	40	51	22	27	33	48	32

٤-٢- فرموله کردن مساله و مدل آنتروپی واژه‌ها

در رویکرد پیشنهادی فرض کنید متغیر تصادفی X نشان‌دهنده تعداد تکرار واژه‌های پرسجوی کاربر در یک سند می‌باشد که آن را به صورت رابطه (۴) نشان می‌دهیم:

$$X = \{x_1 , x_2 , \dots , x_n\} \quad (۴)$$

احتمال رخداد هر یک از واژه‌های پرسجو در اسناد را با رابطه (۵) نشان می‌دهیم:

$$p(X) = \{ p(x_1) , p(x_2) , \dots, p(x_n) \} \quad (۵)$$

که توسط رابطه (۶) محاسبه می‌شود و به این ترتیب مقادیر متغیر تصادفی X نرمال‌سازی می‌شود:

$$p(x_i) = \frac{xi}{\sum_{i=1}^{n} xi} \quad (۶)$$

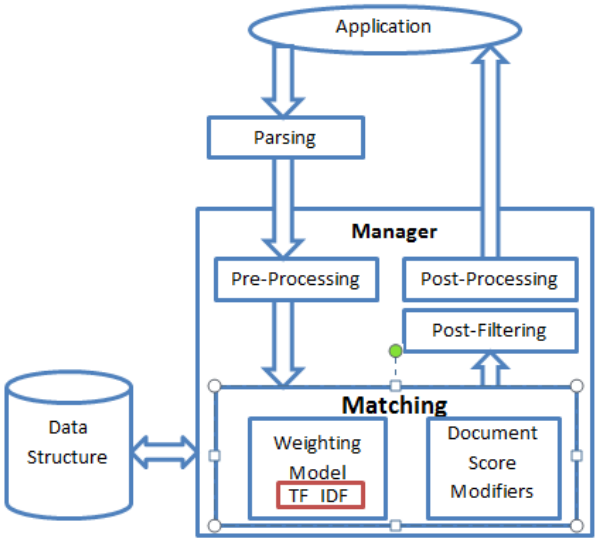
در این رابطه n برابر با تعداد واژه‌های پرسجو می‌باشد. برای این متغیر تصادفی تابع (۷) را نسبت دهیم:

$$\text{Entropy } (X) = - \sum_{i=1}^{n} p(xi) \log_2 p(xi) \quad (۷)$$

نتایج آزمایش‌ها حاکی از آن است که وقتی تعداد تکرار یک واژه پرسش در چند سند با هم برابر می‌شود، این پارامتر نقش تعیین کننده ای در رتبه بندی سندهای مربوطه ایفا می‌کند.

٤-٣- اعمال رویکرد پیشنهادی به معماری موتور جستجو

در این تحقیق برای انجام آزمایش‌ها از سیستم جستجوی Terrier استفاده شده است [24]. بخشی از معماری فاز بازیابی در این سیستم جستجو در شکل (۱) آمده است. پرسش پس از دریافت از Application کاربر، تجزیه شده و به واحد مدیریت تحویل داده می‌شود. در این واحد ابتدا پیش پردازش-هایی بر روی واژه‌های پرسش (مانند حذف کلمات پرتکرار و اعمال الگوریتم Porter Stemmer) اعمال شده و سپس ریشه هریک از واژه‌های پرسش به واحد تطابق ارسال می‌شود. در این واحد، مدل وزنی مناسب اعمال می‌شود و بر اساس آن رتبه هریک از اسناد بررسی شده و رتبه اسناد برای هر پرسش محاسبه می‌شود. بنابراین فاکتور آنتروپی می‌تواند در همین واحد به مدل وزنی TF_IDF اعمال شود تا بر اساس آن رتبه اسناد مربوط به هر پرسش مشخص شود.



شکل (۱): معماری بخش بازیابی موتور جستجوی Terrier

برای اعمال این پارامتر در این مدل، ما این پارامتر را به عنوان یک وزن اضافی به آن اعمال می‌کنیم.به این ترتیب که واژه‌هایی که آنتروپی کمتری داشته باشند،وزن بیشتری را می‌گیرند و این باعث می‌شود که در مجموع اسنادی که آنتروپی کمتری دارند، وزن بیشتری را دریافت کنند.

٥- پارامترهای ارزیابی کارایی

برای ارزیابی کارایی مدل پارامترهای زیر را در نظر می‌گیریم:

۱. R-Precision که به صورت رابطه (۸) تعریف می‌شود وعبارتست از نسبت بین همه اسناد مرتبط بازیابی شده به تعداد کل اسناد مرتبطی که در کلکسیون وجود دارد.

R-Precision = (r/R) (۸)

r تعداد اسناد مرتبط بازیابی شده و R تعداد کل اسناد مرتبط موجود درکلکسیون است.

۲. Precision@k که دقت سیستم جستجورا بعد از k سند بازیابی شده برای هر پرسش محاسبه می‌کند.

۳. Average Precision که با رابطه (۹) محاسبه می شود:

Average Precision = (۹)

$$\frac{\sum_{k=1}^{n} P(k)*rel(k)}{numberOfRelevantDocument}$$

که در آن n تعداد کل اسناد بازیابی شده است و P(k) = Precision@k و rel(k) یک تابع شاخص است که اگر kامین سند بازیابی شده، مرتبط باشد دارای مقدار یک می‌باشد، وگرنه صفر می‌باشد.

در این مقاله برای ارزیابی کارایی راهکار پیشنهادی از سه مجموعه تست به زبان انگلیسی استفاده می‌شود که عبارت اند از:

۱. مجموعه CACM که شامل مقالات منتشر شده در مجله ACM بین سال های ۱۹۵۸ تا ۱۹۷۹ بوده و توسط این موسسه منتشر شده است. این مجموعه شامل ۳۲۰۴ سند و ۶۴ پرسش استاندارد می‌باشد.

۲. مجموعه Medline که شامل مجموعه‌ای از مقالات مجلات پزشکی بوده و دارای ۱۰۲۳ سند و۳۰ پرسش استاندارد است.

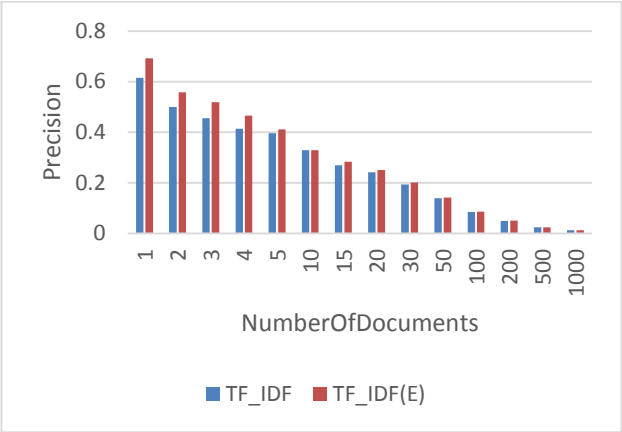
۳. مجموعه CISI که شامل عنوان، خلاصه و نویسندگان مجموعه مقالات برتری می‌باشد که در علوم کامپیوتر بین سال‌های ۱۹۶۹ تا ۱۹۷۷ از بقیه مقالات بیشتر مورد ارجاع قرار گرفته‌اند. این مجموعه شامل۱۴۶۰ سند و ۱۱۲ پرسش استاندارد می‌باشد.

برای ایندکس‌گذاری، بازیابی و ارزیابی نتایج از موتور جستجوی Terrier نسخه ۴.۱ استفاده می‌شود.

۵-۱- نتایج ارزیابی

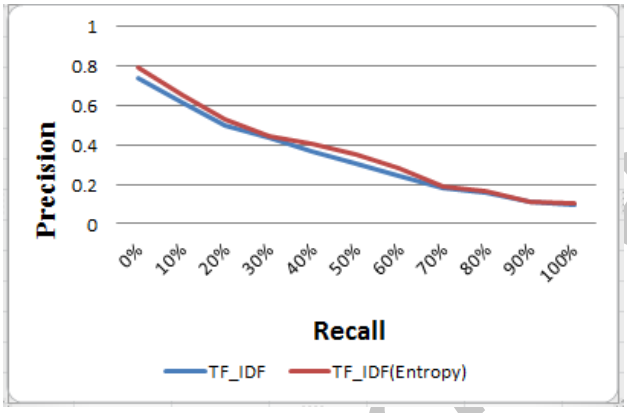
برای ارزیابی تاثیر این پارامتر، آن را به مدل TF_IDF نرم‌افزار Terrier اضافه نموده و نتایج حاصل از ارزیابی کلکسیون‌های تست با مدل اصلی TF_IDF مقایسه شده است.

در شکل (۲) نتایج حاصل از ارزیابی دقت بازیابی اسناد در مدل TF_IDF با اعمال این پارامتر بر روی کلکسیون CACM نشان داده شده است.



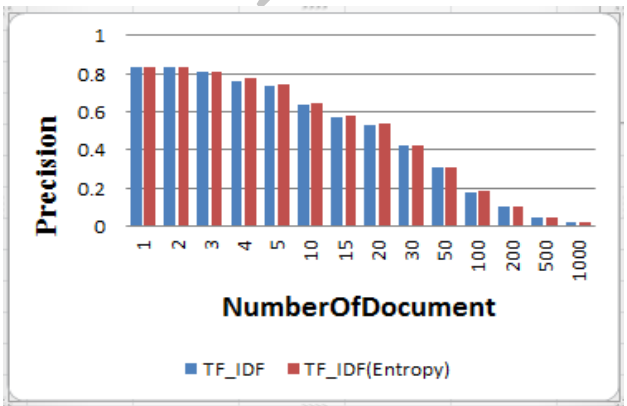
شکل (۲): مقایسه دقت (p@k) حاصل از بازیابی اسناد در موتور Terrier با الگوریتم TF_IDF و TF_IDF(E) با مجموعه CACM

در شکل (۳) نمودار Precision-Recall مربوط به این مدل، قبل و بعد از اعمال پارامتر آنتروپی در کلکسیون CACM نشان داده شده است.



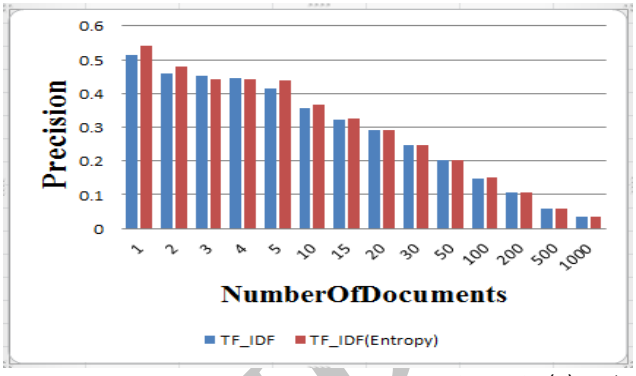
شکل (۳): نمودار Precision-Recall مدلTF_IDF قبل و بعد از اعمال پارامتر آنتروپی در CACM

در شکل (۴) نتایج حاصل از ارزیابی دقت بازیابی اسناد در مدل TF_IDF با اعمال این پارامتر بر روی کلکسیون Medline نشان داده شده است:



شکل (٤): مقایسه دقت (p@k) حاصل از بازیابی اسناد در موتور Terrier با الگوریتم TF_IDF و TF_IDF(E) با مجموعه Medline

در شکل (۵) نتایج حاصل از ارزیابی دقت بازیابی اسناد در مدل TF_IDF با اعمال این پارامتر بر روی کلکسیون CISI نشان داده شده است:



شکل (۵): مقایسه دقت (p@k) حاصل از بازیابی اسناد در موتور Terrier با الگوریتم TF_IDF و TF_IDF(E) با مجموعه CISI

٦-نتیجه گیری

در این پژوهش پارامتر آنتروپی تعداد تکرار واژه های پرسش در اسناد را به عنوان پارامتر موثر در بهبود دقت الگوریتم‌های رتبه بندی، ارائه کردیم. به این ترتیب با اعمال این پارامتر به مدل وزنی TF_IDF، مدل وزنی TF_IDF(E) معرفی شد. با در نظر گرفتن این پارامتر می‌توانیم این چالش را مطرح کنیم که هر چه توزیع واژه‌های پرسش در یک سند منظم‌تر باشد، با در نظر گرفتن سایر پارامترهای موثر، آن سند می‌تواند رتبه بالاتری را در نتایج جستجو به خود اختصاص دهد.

نتایج مربوط به بهبود دقت مربوطه حاصل از ارزیابی اسناد کلکسیون‌های مختلف با اعمال این پارامتر به مدل وزنی TF_IDF در جدول (۳) نشان داده شده است:

جدول (۳): درصد افزایش دقت بازیابی کلکسیون‌ها با اعمال پارامتر آنتروپی

کلکسیون	اندازه	درصد افزایش دقت a	درصد افزایش دقت b
CACM	۲.۲ مگابایت	٪۱۶	٪۱۱
CISI	۲.۱ مگابایت	٪۱۲	٪۸
Medline	۱.۱ مگابایت	٪۶	٪۴

a: R Precision
b: Average Precision

نتایج تست حاکی از آن است که با افزایش سایز مجموعه تست، تاثیر این پارامتر بیشتر دیده می‌شود.

در آینده می‌توان این پارامتر را به سایر مدل‌های وزنی نیز اعمال کرد و نتایج تاثیر آنها را با هم مقایسه کرد. نتایج پژوهش حاکی از آن است که پارامتر آنتروپی تکرار واژه‌های پرسش، اهمیت زیادی داشته و همواره می‌توان آن را به صورت عاملی در نظر گرفت که باید با تکرار واژه‌های پرسش موازنه داشته باشد، تا به این ترتیب بتوان رتبه بالاتری را برای یک سند در نظر گرفت.

در آینده باید روی این پارامتر مطالعات بیشتری صورت بگیرد و با اعمال پاره ای تغییرات در آن و اعمال در مدل رتبه‌بندی، می‌توان میزان تاثیرگذاری آن را افزایش داد. در آینده می‌توان این پارامتر را به سایر مدل‌های وزنی نیز اعمال کرد و نتایج تاثیر آنها را با هم مقایسه کرد.

مراجع

[1] Barsagade, Naresh, Kaufman, Morgan,"Web Usage Mining And Pattern Discovery: A Survey Paper", on Machine Learning, pages 167-174. CSE 8331, Dec, 2003.
[2] Duhan, Neelam, Sharma, A.K, Bhatia, Komal, Kumar," Page Ranking Algorithms: A Survey", 2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.
[3] Malhotra, Dheerja, Malhotra, Monica, Rishi O.P., "An Innovative Approach of Web Page Ranking Using Hadoop- and Map Reduce-Based Cloud Framework", In: Aggarwal V., Bhatnagar V., Mishra D. (eds) Big Data Analytics. Advances in Intelligent Systems and Computing, vol 654. Springer, Singapore, 2018.
[4] Pardakhe, N.V, Keole, Prof.R., "Analysis of Various Web Page Ranking Algorithms in Web Structure Mining", International Journal of Advanced Research in Computer and Communication Engineering, Vol.2, Issue 12, December 2013.
[5] Gomes da, Costa Miguel, Gong, Júnior Zhiguo," Web Structure Mining: An Introduction", Proceedings of the 2005 IEEE International Conference on Information Acquisition June 27 - July 3, 2005.
[6] Shinde, Mayuri, Girase, Sheetal, "A Survey of various Web Page Ranking Algorithms", International Journal of Computer Applications (0975 – 8887), Volume 132 – No.10, December 2015.
[7] Kumar, Kaushal, Fungayi, Abhaya, Mukoko, Donewell, "PageRank algorithm and its variations: A Survey report", OSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727 Volume 14, Issue 1, PP 38-45, Sep. - Oct 2013.
[8] Xing, W., Ghorbani, A., "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04), IEEE, pp. 305- 314, 2004.
[9] Salton, G., Buckley, C., "Weighting Approaches in Automatic Text Retrieval". In Information Processing and Management, Vol. 24, No.5, pp. 513-523, 1998.
[10] Pokorny, Jaroslav, Smizansky, Jozef, "Page Content Rank: An Approach to the Web Content Mining", IADIS International Conference, 2005.
[11] Kleinberg, J., "Authorative Sources in a Hyperlinked Environment", Proceedings of the 23rd annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.
[12] Cohn, D., Chang, H., "Learning to Prob abilistically identify Authoritative Documents". In Proceedings of 17th International Conf, A on Machine Learning, pages167-174. Morgan Kaufmann, San Francisco, CA, 2000.
[13] Li, Longzhuang, Shang, Yi, Zhang, Wei, "Improvement of HITS- based Algorithms on Web Documents",

WWW2002, May 7-11, Honolulu, Hawaii, USA. ACM 1-58113-449-5/02/0005, 2002.

[14] Au, Yeung, Ching-man, G. Noll, Michael, Gibbins, Nicholas, Meine, Christoph l, Shadbolt, Nigel, " SPEAR: Spamming-Resistant Expertise Analysis and Ranking in Collaborative Tagging System", Computational Intelligence, Volume 99, Number 000, 2009 .

[15] Kumar, G., Duhan, N., Sharma, A.K., "Page Ranking Based on Number of Visits of Links of Web Page",International Conference on Computer & Communication Technology (ICCCT)-2011, IEEE, pp.11-14, 2011.

[16] Singh, Amar, Sharma Sanjeev, " Role of Page ranking algorithm in Searching the Web: A Survey ", International Journal of Engineering & Technology, Management and Applied Sciences, Vol 1, Issue 1, June 2014.

[17] Li, C., Han, J., He, G., Jin, X., Sun, Y., Yu, Y., Wu, T., " Fast Computation of SimRank for Static and Dynamic Information Networks", Published in ACM, Print ISBN No: 978-1-60558-9045-9, on 22-26 March 2010.

[18] Gupta, Dr. D., Singh, D., "User Preference Based Page Ranking Algorithm", International Conference on Computing, Communication and Automation (ICCCA2016), ISBN: 978-1-5090-1666-2/16/$31.00 ©2016 IEEE, 2016.

[19] Chong, T., "A Kind of Algorithm For Page Ranking Based on Classified Tree In Search Engine", International Conference on Computer Application and System Modeling (ICCASM), IEEE, pp.538-541, 2010.

[20] Nagappan , V.K, Elango, Dr. P. ," Agent Based Weighted Page Ranking Algorithm for Web Content Information Retrieval", 978-1-4799-7623-2/15/$31.00_c 2015 IEEE.

[21] Prajapat, R.i, Kuma, S.r, "Enhanced Weighted PageRank Algorithm Based on Content and Link Visits", 978-9-3805-4421-2/16/$31.00_(C) 2016 IEEE.

[22] http://www.terrier.org/.

[23] Breitinger, Corinna, Gipp, Bela, Langer, Stefan (2015-07-26). "Research-paper recommender systems: a literature survey". International Journal on Digital Libraries. 17 (4): 305–338. doi:10.1007/s00799-015-0156-0. ISSN 1432-5012, 2015.

زیرنویس

[1] Web Mining
[2] Web Structure Mining
[3] Web Content Mining
[4] Web Usage Mining
[5] Similarity
[6] Term Weighting
[7] Recommender Systems
[8] Judgment