دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

عنوان:

بهبود بازدهی موتورهای جستجو با تکیه بر تکنیکهای تحلیلی گراف وب

نگارش: پدرام قدس نیا

استاد راهنما: دکتر ناصر یزدانی

پایان‌نامه برای دریافت درجهٔ کارشناسی ارشد در رشتهٔ

مهندسی نرم‌افزار

اسفند ۱۳۸۶

دانشگاه تهران

به نام خدا

دانشگاه تهران

معاونت تحصیلات تکمیلی
پردیس دانشکده های فنی
دانشگاه تهران

پردیس دانشکده های فنی

دانشکده مهندسی برق و کامپیوتر

گواهی دفاع از پایان نامه کارشناسی ارشد

هیأت داوران پایان نامه کارشناسی ارشد آقا/خانم پدرام قدس نیا در رشته مهندسی برق و کامپیوتر، گرایش : نرم افزار

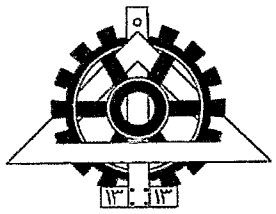
با عنوان: " بهبود بازدهی موتورهای جستجو با تکیه بر تکنیکهای تحلیلی گراف وب "

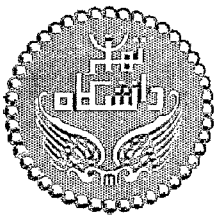
در تاریخ ۱۳۸۶/۱۲/۱۳ نمره نهایی پایان نامه: به عدد ۲۰ به حروف بیست

و درجه عالی ارزیابی نمود.

مشخصات هیأت داوران	نام و نام خانوادگی	مرتبه دانشگاهی	دانشگاه یا موسسه	امضاء
۱-استاد راهنما استاد راهنمای دوم (حسب مورد)	دکتر ناصر یزدانی	دانشیار	تهران	
۲-استاد مشاور	---	---	---	
۳ -استاد مدعو خارجی (یا استاد مشاور دوم)	دکتر حسن ابوالحسنی	استادیار	شریف	
۴-استاد مدعو داخلی	دکتر بهزاد مشیری	استاد	تهران	
۵-داور و نماینده کمیته تحصیلات تکمیلی دانشکده	دکتر فتانه تقی یاره	استادیار	تهران	

تذکر:این برگه پس از تکمیل توسط هیأت داوران در نخستین صفحه پایان نامه درج می گردد.





دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

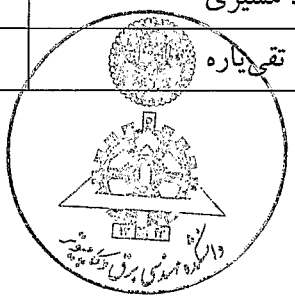
پایان نامه برای دریافت درجه کارشناسی ارشد در رشته مهندسی نرم‌افزار

بهبود بازدهی موتورهای جستجو با تکیه بر تکنیکهای تحلیلی گراف وب

نگارش پدرام قدس نیا

این پایان نامه در تاریخ ۱۳۸۶/۱۲/۱۳ در مقابل هیات داوران دفاع گردید و مورد تصویب قرار گرفت.

معاونت آموزشی تحصیلات تکمیلی پردیس دانشکده‌های فنی: دکتر جواد فیض	
رئیس دانشکده مهندسی برق و کامپیوتر: دکتر پرویز جبه‌دار مارالانی	
معاونت پژوهشی و تحصیلات تکمیلی دانشکده: دکتر سعید نادر اصفهانی	
استاد راهنما: دکتر ناصر یزدانی	
عضو هیأت داوران: دکتر حسن ابوالحسنی	
عضو هیأت داوران: دکتر بهزاد مشیری	
عضو هیأت داوران: دکتر فتانه تقی‌یاره	



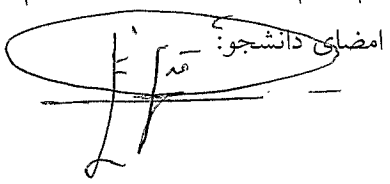
تعهد نامه اصالت اثر

اینجانب **پدرام قدس نیا** تأیید می‌نمایم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب است و به دستاوردهای پژوهشی دیگران که در این نوشته از آنها استفاده شده است مطابق مقررات ارجاع گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم سطح یا بالاتر ارائه نشده است.

کلیه حقوق مادی و معنوی این اثر متعلق به پردیس دانشکده‌های فنی دانشگاه تهران می باشد.

نام و نام خانوادگی دانشجو: **پدرام قدس نیا**

امضای دانشجو:



تقدیم به پدرم

که به من اندیشیدن را آموخت

با تشکر از استاد راهنمای عزیزم جناب آقای دکتر یزدانی که رهنمودهای ارزنده‌ی ایشان در نگارش این پایان نامه برایم همواره روشنگر و چاره ساز بود.

با تقدیر فراوان از دوست گرامیم جناب آقای مهندس علی‌محمد زارع بیدکی که بی‌شک بدون مساعدت و راهنمایی ایشان پژوهشهای انجام شده در این پایان نامه کیفیت لازم را نمی‌داشت.

با سپاس از دوستان و همکاران عزیزم آقایان محمد کیهانی، فرید امیرغیاثوند، پویا توکلی و بنیامین خداوردی که مرا در به اتمام رسانیدن این پایان‌نامه یاری بسیار رساندند.

با تشکر از همسر نازنینم که در ویرایش این پایان نامه مرا یاری کرد و با سپاس از پدر، مادر و خواهر گرامیم و سایر عزیزانی که در طول نگارش این پایان نامه زحمات زیادی برای من کشیدند.

چکیده

از زمان پیدایش شبکه‌ی جهانی اینترنت تا کنون حجم وب همواره رو به افزایش بوده است. استفاده از گستره‌ی اطلاعاتی وب یکی از نیازهای روزمره انسانها در جوامع بشری امروزی را تشکیل می‌دهد. علاوه بر اینکه عدم وجود نظارت و کنترل بر نحوه‌ی تولید محتوا در وب، منجر به تولید اطلاعات فراوانی با سطوح کیفیت، اهمیت و درستی متفاوت شده، استفاده‌کنندگان از وب را نیز قشر وسیعی از کاربران با سلایق، علایق، نقطه نظرات، تحصیلات، توانایی‌ها و سطوح فرهنگی متنوع تشکیل می‌دهند.

جستجو در این بازار آشفته و یافتن نتایجی با کیفیت و مطابق با نظر کاربرانی با این سطح از تنوع در طرز فکر، به یکی از چالش برانگیزترین مسائل تحقیقاتی در حوزه‌ی بازیابی اطلاعات تبدیل شده‌است. یکی از مهمترین چالشها در این میان، چگونگی رتبه‌بندی نتایجی است که موتورهای جستجو در پاسخ به پرس‌وجوی کاربران پیدا می‌کنند. رضایت کاربران از موتور جستجو بستگی مستقیم به مرتبط بودن و با کیفیت بودن نتایجی دارد که در ابتدای لیست نتایج در پاسخ به پرس‌وجویشان به نمایش در می‌آید.

در این پژوهش مدلها و الگوریتمهای جدیدی ارائه شده است که در سه حوزه، کاهش مشکلات مختلفی را که در رابطه با رتبه‌بندی نتایج جستجو با آنها روبرو هستیم هدف قرار می‌دهند. در بخش اول نسخه‌ی جدیدی از الگوریتم معروف PageRank که در موتور جستجوی موفق Google مورد استفاده قرار گرفته ارائه شده است. این مدل جدید بدون تحمیل پیچیدگی زمانی و هزینه‌ی حافظه‌ی اضافی، با نزدیک‌تر کردن مدل ریاضی نهفته در پس الگوریتم PageRank به واقعیت، به بهبود کیفیت نتایج جستجو کمک می‌کند. در بخش بعد روش رتبه‌بندی متفاوتی ارائه

گردیده که با بکارگیری نظام تنبیه و پاداش سعی دارد تا تاثیر مشکل غنی تر شدن اغنیاء را در رتبه‌بندی نتایج جستجو کاهش دهد. در نهایت الگوریتمی تطبیق‌پذیر[1] برای تجمیع معیارهای رتبه‌بندی مبتنی بر *محتوا*[2] و مبتنی بر *ساختار گراف وب*[3] ارائه شده است. این الگوریتم با استفاده از یادگیری تقویتی، از رفتار کاربران موتور جستجو در مواجهه با نتایج جستجوهای قبلی یاد می‌گیرد که چگونه نحوه‌ی تجمیع را اصلاح کند تا سطح بالاتری از رضایت کاربران را در جستجوهای بعدی به همراه داشته باشد.

[1] Adaptive
[2] Content-based
[3] Connctivity-based

فهرست مطالب

فهرست شکل‌ها

فصل ۱. مقدمه

با توجه به رشد روز افزون حجم وب و افزایش چشمگیر تعداد صفحات موجود در آن و با درنظر گرفتن اینکه هم‌اکنون در عصر اطلاعات به سر میبریم و در زمان حاضر، نیاز به در دسترس بودن سریع اطلاعات درست و با کیفیت، بخشی از نیازهای زندگی روزمره ما را تشکیل می‌دهد، بهبود و بهینه سازی ابزارهای جستجو، به عنوان اهرمهای قدرتمندی در جهت تسریع و تسهیل این امر، از اهمیت فوق العاده ای برخوردار می‌باشد. در حال حاضر حتی قوی ترین موتورهای جستجوی جهان تنها موفق به تحت پوشش قرار دادن بخش کوچکی از حجم عظیم وب شده‌اند و در بسیاری از موارد در مورد همان بخش کوچک نیز نتایج جستجوئی با کیفیت مورد نظر ارائه نمی‌کنند.

پیشرفتهای اساسی در تکنولوژی موتورهای جستجو را می‌توان به نیمه‌ی دوم دهه‌ی نود میلادی نسبت داد. در ابتدای امر، عمده‌ی پیشرفتهای حاصله در این زمینه مدیون نتایج تحقیقات انجام شده در زمینه بازیابی اطلاعات به *روشهای سنتی*[4] بود [۱] که این تحقیقات سالها قبل از به وجود آمدن وب آغاز شده بودند. در بازیابی اطلاعات به روشهای سنتی، عمدتاً بر اساس محاسبه‌ی شباهتِ مابین *پرس‌وجوی*[5] کاربر و اسناد بازیابی شونده عمل می‌شود.

با این حال وب دارای خصوصیتهایی است که موجب می‌شود بکارگیری صِرف روشهای سنتی بازیابی اطلاعات، نتایج مناسبی را به دنبال نداشته باشند. در مقایسه با مجموعه‌هایی از اطلاعات که در روشهای سنتی بازیابی اطلاعات مطرح بودند، در وب ما با تعداد بسیار بیشتری مستندات قابل بازیابی روبرو هستیم که این مستندات علاوه بر اینکه هر روز در حال افزایش هستند، حوزه‌ی بسیار وسیع‌تری از اطلاعات را نیز تحت پوشش قرار می‌دهند. محتوای صفحات وب دائماً در حال تغییر است و برای تولید محتوا در وب اگرچه استانداردهایی تدوین شده، ولی در کل و در مقایسه با بازیابی اطلاعات سنتی، در اینجا کنترل و حاکمیتی قوی بر نحوه‌ی تولید محتوا وجود ندارد. این آزادی عمل در نحوه‌ی تولید محتوا طبیعتاً منجر به افزایش تعداد تولیدکنندگان محتوا می‌گردد. تولید کنندگانی که ذاتاً از افکار، عقاید، سلایق و توانمندی های متفاوتی برخوردار می‌باشند. در نتیجه در وب با انبوهی از اطلاعات ناهمگن و بعضاً متناقض روبرو خواهیم بود که تشخیص درست از

[4] Traditional Information Retrieval
[5] Query

نادرست، متناسب از نامتناسب، مرتبط از نامرتبط و باکیفیت از بی‌کیفیت در این کلاف سر در گم خود مقوله‌ای قابل تامّل است. بعلاوه کاربران موتورهای جستجو در وب انتظار دارند تا با پرس‌وجوهایی به مراتب کوتاهتر (به طور میانگین ۲.۴ کلمه[۲]) به اطلاعات مورد نظر خود دست یابند و این در حالی است که وب دامنه‌ی بسیار وسیعتری از لغات را آن هم به زبانهای مختلف شامل می‌شود. تفاوتهای مذکور، چالشهای بسیاری را بر سر راه محققان فعال در حوزه‌ی موتورهای جستجو قرار داده‌است که در حال حاضر نیز بسیاری از این چالشها همچنان بر قوت خود باقی هستند.

چالشهای اخیر، دانشمندان فعال در این حوزه را بر آن داشت که در کنار روشهای سنتی بازیابی اطلاعات، به دنبال روشهای موثرتری بگردند تا بتوانند با تکیه بر این روشها از عهده‌ی پاسخگویی به نیاز اطلاعاتی انبوه کاربران رو به رشد شبکه جهانی اینترنت بر آمده و کیفیت نتایج جستجو را که عموماً با معیارهایی چون سطح رضایت کاربر از نتایج و یا میزان ارتباط نتایج یافت‌شده با پرس‌وجوی کاربر بیان می‌شود، بالا ببرند.

یکی از مطرح‌ترین روشهای بکارگرفته شده در سالهای اخیر استفاده از ساختار گراف وب در رتبه‌بندی نتایج جستجو می‌باشد [۱، ۳-۵]. وقتی در صفحه‌ی p پیوندی به صفحه‌ی q وجود دارد، این پیوند به منزله‌ی اطلاعی در مورد صفحه‌ی q است که در صفحه‌ی p وجود دارد. در نتیجه، از ساختار پیوندهای مابین صفحات وب می‌توان به عنوان منبع اطلاعاتی ارشمندی برای بهبود نتایج جستجو استفاده‌کرد. بهره‌بردن از این اطلاعات ارزشمند موجب پیشرفت قابل توجهی در بهبود نتایج موتورهای جستجوی قدرتمندی همچون گوگل [۶] شده‌است.

یکی دیگر از منابع ارزشمند اطلاعاتی برای بهبود کیفیت نتایج جستجو در سالهای اخیر، استفاده از اطلاعات نگهداری شده از سابقه‌ی چگونگی استفاده‌ی کاربران از نتایج جستجو بوده‌است. به این اطلاعات اصطلاحاً اطلاعات *کلیک از گذر داده*[۶] گفته می‌شود[۷-۹]. وقتی در پاسخ به یک پرس‌وجو لیستی از نتایج به کاربر نشان داده می‌شود، کاربر به بررسی نتایج می‌پردازد و با توجه به اطلاعات مختصری که در مورد هر نتیجه در لیست وجود دارد در مورد کلیک کردن و یا کلیک

[6] Click-through Data

نکردن روی هر نتیجه تصمیم می‌گیرد. در نتیجه، کلیک کردن بر روی یک یا چند نتیجه می‌تواند به احتمال زیاد به منزله‌ی ترجیح دادن صفحات کلیک شده بر صفحات کلیک نشده از جانب کاربر تلقی شود. این ترجیحات را که به صورت ضمنی و نه صریح، از رفتار کاربر استخراج می‌شوند اصطلاحاً *بازخورد ضمنی*[7] می‌نامند و محققان با تکیه بر روشها و تکنیکهای مختلفی که عمدتاً بر *یادگیری ماشینی*[8] استوار هستند، در جهت بهبود کیفیت نتایج جستجو از این بازخوردها بهره می‌برند.

علاوه بر اینکه روشهای سنتی بازیابی اطلاعات به تنهایی از پس چالشهایی که در محیط وب با آنها روبرو هستیم بر نمی‌آیند، روشهای مدرن مذکور نیز به تنهایی پاسخگوی حل این معمای پیچیده نمی‌باشند. لذا عموماً برای رتبه‌بندی نتایج، از روشهای ترکیبی استفاده می‌شود. بدین منظور، رتبه‌بندی ارائه شده توسط روشهای مختلف را با یکدیگر تجمیع[9] می‌نماییم و از ترکیب حاصل به یک رتبه‌بندی جدید دست خواهیم یافت. برای انجام عمل تجمیع تا کنون فعالیتهای تحقیقاتی بسیار زیادی انجام شده‌است. با این حال ماهیت ناهمگون و در حال تغییر وب، ذاتاً ایجاب می‌کند که هیچ روشی همیشه بهترین روش نباشد. لذا تحقیقات در این راستا همچنان ادامه دارد و در این میان روشهایی که به نحوی توانایی تطابق دائمی با طبیعت متغییر محتوا و ساختار وب را داشته باشند بیشتر مورد توجه خواهند بود.

۱-۱ فعالیت های صورت گرفته

در پژوهش اخیر سعی بر آن بوده‌است که مساله‌ی رتبه‌بندی صفحات وب با تکیه بر تکنولوژی‌ها و دستاوردهای بازیابی اطلاعات مدرن در سالهای اخیر مورد بازنگری قرار گیرد و مدلهای جدیدی برای استفاده‌ی بهتر از اطلاعات موجود، در راستای بهبود نتایج جستجو ارائه گردد. کارهای انجام شده در این تحقیق را می‌توان در سه حوزه‌ی زیر خلاصه نمود :

۱- ارائه‌ی نسخه‌ی جدیدی از الگوریتم PageRank [۳، ۱۰، ۱۱] (که به عنوان یکی از معیارهای مهم برای رتبه‌بندی صفحات وب توسط موتور جستجوی قدرتمند گوگل مورد

[7] Implicit Feadback
[8] Machine Learning
[9] Aggregation

استفاده قرار گرفته است). در الگوریتم PageRank هر پیوندی که از صفحه‌ای به صفحه‌ی دیگری در وب اشاره می‌کند به منزله‌ی توصیه‌ی تولیدکننده‌ی صفحه‌ی مبدا در مورد اهمیت صفحه‌ی مقصد تلقی می‌شود. در الگوریتم جدید که آن را **WPR**[10] نامیده‌ایم علاوه بر توصیه‌ی سازندگان صفحات، میانگین توصیه‌های بازدیدکنندگان صفحات نیز در سنجش میزان اهمیت آنها نقش خواهد داشت. WPR بدون افزودن هزینه‌ی حافظه و پیچیدگی محاسباتی اضافی محسوسی، جای خالی نظر بازدیدکنندگان صفحات در الگوریتم PageRank را پر می‌کند.

۲- ارائه‌ی الگوریتم جدیدی که با تکیه بر مفهوم دموکراسی و با استفاده از سیستم پاداش و تنبیه سعی دارد تا مشکل *غنی‌تر شدن اغنیا*[11] [۱۲، ۱۳] را که یکی از چالشهای مطرح در الگوریتمهای رتبه‌بندی مبتنی بر ساختار گراف وب به شمار می‌رود، تقلیل بخشد. در الگوریتمهای رتبه‌بندی مبتنی بر ساختار گراف وب، صفحاتی که از محبوبیت عام برخوردار می‌شوند، مرتباً بر محبوبیتشان افزوده می‌شود. این امر فرصت مطرح شدن را از صفحات با اهمیتی که تازه متولد شده‌اند باز می‌ستاند و در نتیجه مدت زمان بسیار زیادی به طول خواهد انجامید تا صفحاتِ با اهمیتِ جدید و تازه متولد شده بتوانند به رتبه‌های بالا در نتایج جستجو راه پیدا کنند. همچنین روش جالبی برای ترکیب نتایج حاصل از این الگوریتم با الگوریتم PageRank ارائه شده‌است.

۳- ارائه الگوریتمی *سازگار*[12] برای تجمیع معیارهای رتبه‌بندی *مبتنی بر محتوا*[2] و *مبتنی بر ساختار گراف وب*[3] با استفاده از اطلاعات کلیک از گذر داده برای بهبود کیفیت نتایج جستجو. این الگوریتم که آن را **A3CRank**[13] نامیده‌ایم، از عملگر ریاضی تجمیع OWA برای تجمیع نتایج حاصل از الگوریتمهای رتبه‌بندی بهره می‌برد. در این الگوریتم به کمک یادگیری تقویتی، نحوه‌ی تجمیع از روی رفتار کاربران در مواجهه با نتایج پرس‌وجوهای قبلی

[10] Weighted PageRank
[11] Rich-get-richer Problem
[12] Adaptive
[13] Adaptive method based on **Content**, **Connectivity** and **Clickthrough** data **Rank**

یادگرفته می‌شود. ایده‌ی اولیه‌ی این متد از روش مشابهی که در *ابرموتور جستجو*[14] برای رتبه‌بندی مجدد نتایجِ دریافت شده از موتورهای جستجوی زیرین استفاده شده [۱۴-۱۶] گرفته شده‌است، ولی در اینجا به جای موتورهای جستجوی زیرینِ ابرموتور جستجو، الگوریتمهای مختلف رتبه‌بندی را پیش روی داریم.

۱-۲ ساختار کلی پایان نامه

این پایان نامه به هفت فصل تقسیم شده است. در فصل دوم در ابتدا به معرفی اجمالی اجزای اصلی موتورهای جستجو خواهیم پرداخت و خلاصه‌ای از زمینه‌های تحقیقاتی، کارهای انجام شده و چالشهای پیش روی در مورد هر بخش را ارائه خواهیم نمود. سپس حوزه‌ی مطالعاتی مورد نظر در این پایان‌نامه را که بهبود کیفیت نتایج جستجو از نقطه نظر رتبه‌بندی نتایج جستجو می‌باشد مرور خواهیم کرد. در فصل سوم، ابتدا اصول و تعاریف اولیه‌ی گراف وب و خواص آن و همچنین روابط ریاضی مربوطه را بیان خواهیم نمود. در فصل چهارم، الگوریتم WPR را به عنوان نسخه‌ی بهبود یافته‌ای از الگوریتم PageRank ارائه خواهیم نمود. در الگوریتم PageRank تنها توصیه‌ی سازندگان صفحات وب در مورد اهمیت سایر صفحات در رتبه‌بندی دخیل است در حالی که WPR می‌کوشد تا جای خالی میانگین توصیه و نظر بازدیدکنندگانِ از صفحات را نیز بدون تحمیل پیچیدگی محاسباتی و هزینه‌ی حافظه‌ی اضافی قابل توجهی در رتبه‌بندی بکار گیرد. در فصل پنجم، با استفاده از یک نظام مبتنی بر تنبیه و پاداش مدل جدیدی برای رتبه‌بندی مستقل از پرس‌وجو ارائه خواهیم داد و نشان می‌دهیم که استفاده از این مدل در تقلیل مشکل غنی‌تر شدن اغنیاء تاثیر بسزایی دارد. در فصل ششم با الهام از یکی از روشهای تجمیع اطلاعات در ابر موتورهای جستجو، راهکاری سازگار برای تجمیع معیارهای رتبه‌بندی مبتنی بر محتوا و معیارهای رتبه‌بندی مبتنی بر ساختار گراف وب با استفاده از اطلاعات کلیک از گذر داده ارائه خواهیم نمود و در نهایت در فصل هفتم به جمع‌بندی تحقیقات صورت گرفته و نتایج به‌دست‌آمده خواهیم پرداخت و رهنمودهایی برای ادامه تحقیقات در این زمینه ارائه خواهیم کرد.

[14] Meta Search Engine