دانشکده فني و مهندسي
گروه علمي مهندسي کامپيوتر و فناوري اطلاعات
پايان‌نامه کارشناسي ارشد
رشته مهندسي کامپيوتر – گرايش نرم‌افزار

الگوریتم بهینه Plane sweep با استفاده از داده‌های تکراری موجود در سند

نگارش:
الهه مقیمی هنجنی

استاد راهنما:
دکتر مهدی جوانمرد

استاد مشاور:
دکتر واهه آغازاریان

آذر ۱۳۹۱

دانشکده فني و مهندسي
گروه علمي مهندسي کامپيوتر و فناوري اطلاعات
پايان‌نامه کارشناسي ارشد
رشته مهندسي کامپيوتر – گرايش نرم‌افزار

الگوریتم بهینه Plane sweep با استفاده از داده‌های تکراری موجود در سند

نگارش:
الهه مقیمی هنجنی

استاد راهنما:
دکتر مهدی جوانمرد

استاد مشاور:
دکتر واهه آغازاریان

آذر ۱۳۹۱

جمهوری اسلامی ایران
وزارت علوم تحقیقات و فناوری

مرکزی

تاریخ .....
شماره .....
پیوست .....

صورتجلسه دفاع از پایان نامه کارشناسی ارشد

دانشگاه پیام نور استان تهران

جلسه دفاع از پایان نامه کارشناسی ارشد الهه مقیمی هنجنی

رشته مهندسی کامپیوتر (نرم افزار)

تحت عنوان

" الگوریتم بهینه Plane sweep با استفاده از داده های تکراری موجود در سندهای داده ای "

جلسه دفاع با حضور داوران نامبرده ذیل درروز سه شنبه مورخ ۱۳۹۱/۰۹/۰۷ ساعت ۱۱ در مرکز ری برگزار شد و پس از بررسی پایان نامه مذکور با نمره به عدد ۱۸/۵ به حروف هجده و نیم و با درجه عالی مورد قبول واقع شد ☑ نشد ☐

هیات داوران :

داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
استاد راهنما	جناب آقای دکتر مهدی جوانمرد	استادیار	
استاد مشاور	جناب آقای دکتر واهه آغازاریان	استادیار	
استاد داور	جناب آقای دکتر اکبر فرهودی نژاد	استادیار	
نماینده تحصیلات تکمیلی	جناب آقای دکتر اکبر فرهودی نژاد	استادیار	

آیا پایان نامه مذکور نیاز به اصلاحات دارد؟

شهر ری،جاده ورامین، سه راه تقی آباد،انتهای خیابان شهید عربخواری، ساختمان مرکزی شهید دکتر چمران (شماره ۱)
کد پستی: ۱۸۶۵۸۶۵۳۹۹
تلفن: ۳۳۴۱۶۸۱۱-۱۲
سایت دانشگاه: http://teh-rey.pnu.ac.ir
ایمیل دانشگاه info@shahrerey.tpnu.ac.ir

شهر ری،خیابان ابن بابویه ، ابتدای خیابان میر عابدینی، پلاک ۹۰ ساختمان شیخ صدوق (شماره ۲)
کد پستی: ۱۸۶۴۶۶۶۴۸۹
تلفن: ۳۳۳۸۹۹۶۴
۳۳۳۸۹۹۷۰
سایت دانشگاه: http://teh-rey.pnu.ac.ir
ایمیل دانشگاه info@shahrerey.tpnu.ac.ir

اینجانب دانشجوی ورودی سال مقطع کارشناسی‌ارشد رشته
گواهی می‌نمایم چنانچه در پایان‌نامه خود از فکر، ایده و نوشته دیگری بهره گرفته‌ام با نقل قول مستقیم یا غیرمستقیم منبع و ماخذ آن را نیز در جای مناسب ذکر کرده‌ام. بدیهی است مسئولیت تمامی مطالبی که نقل قول دیگران نباشد برعهده خویش می‌دانم و جوابگوی آن خواهم بود.

نام و نام‌خانوادگی دانشجو

تاریخ و امضاء

اینجانب دانشجوی ورودی سال مقطع کارشناسی‌ارشد رشته
گواهی می‌نمایم چنانچه بر اساس مطالب پایان‌نامه خود اقدام به انتشار مقاله، کتاب و ... نمایم ضمن مطلع نمودن استاد راهنما، با نظر ایشان نسبت به نشر مقاله، کتاب و ... و به صورت مشترک و با ذکر نام استاد راهنما مبادرت نمایم.

نام و نام‌خانوادگی دانشجو

تاریخ و امضاء

کلیه حقوق مادی مترتب از نتایج مطالعات، آزمایشات و نوآوری ناشی از تحقیق موضوع این پایان‌نامه متعلق به دانشگاه پیام‌نور می‌باشد.

آذر ۱۳۹۱

تقدیم به

پدر و مادر عزیزم به خاطر زحمات بی دریغشان

تشکر و قدردانی

با سپاس از زحمات استاد محترم راهنما، جناب آقای دکتر جوانمرد و زحمات استاد محترم مشاور جناب آقای دکتر آغازاریان، که همواره برای من الهام‌بخش ایده و دیدگاهی تازه نسبت به موضوع بوده‌اند. همچنین از زحمات جناب آقای دکتر قربان‌نیا نیز، به‌دلیل رهنمودها و تشویق‌های موثرشان، که زمینه‌ساز این تحقیق بوده است، صمیمانه قدردانی می‌کنم.

در انتها از پدر و مادر عزیزم به‌خاطر پشتیبانی همیشگی‌شان در زندگی سپاسگزارم.

آذر ماه ۱۳۹۱

الهه مقیمی

چکیده

در این تحقیق الگوریتم plane sweep اصلاح شده (IPNWPSR)، ارائه شده است. در الگوریتم پیشنهادی با استفاده از پارامترهای موثر، روشی را ارائه دادیم که تعداد کلمات تکراری پشت‌سرهم در سند را شناسایی کرده و همچنین با حذف بازه‌های غیرمفید در جستجو، تعداد مقایسه‌ها را کاهش داده و سرعت الگوریتم را افزایش می‌دهد. در الگوریتم فوق ارتباط میان پرسش و سند براساس فهرستی از آفست‌های کلمه کلیدی مورد جستجو است.

در این رهیافت، راهکاری برای کاهش تعداد مقایسه‌ها با استفاده از شمارنده برای داده‌های تکراری پشت سرهم در جستجو ارائه شده است که باعث شده پیچیدگی زمانی الگوریتم پایه کاهش یابد و در عین حال الگوریتم قابل اعتماد و پایدار بوده و به طور کلی الگوریتمی کارا و بهینه به خصوص در حجم داده بالا خواهیم داشت.

واژه‌های کلیدی: الگوریتمplane sweep ، داده‌های تکراری، بازیابی متن، جستجوی proximity، *جستجوی بخشی.*

فهرست مطالب

عنوان | صفحه

فهرست جدول‌ها

عنوان **صفحه**

فهرست شکل‌ها

عنوان | صفحه

فهرست علائم اختصاری

WPSR	Word Plane Sweep Replicated	الگوریتم Plane Sweep با در نظرگرفتن داده‌های تکراری
IPNWPSR	Iterated Partial Neighbor Word Plane Sweep Replicated	جستجوی بخشی همسایگی با داده‌های تکراری در الگوریتم Plane Sweep
docID	Document ID	شماره سند
HTML	Hypertext Markup Language	زبان نشانه‌گذاری ابرمتن
Zlib	Zip Library	کتابخانه فشرده‌سازی
URL	Universal Resource Locator	سایت جهانی منابع
ISAM	Index Sequential Access Mode	حالت دسترسی ترتیبی اندیس
DocInfo	Document Information	اطلاعات سند
TSPR	Topic Sensitive PageRank	الگوریتم رتبه‌بندی سند
POS	position	موقعیت
D_A	Distance between two A vectors	فاصله میان دو بردار
C_n	Number of compares	تعداد مقایسه‌ها
R	Ratio	نرخ
W_{sim}	Word Similarity	ضریب تشابه کلمات

فصل اول

مقدمه

مقدمه

۱-۱- پیشگفتار

وقتی یک پرسش را مورد جستجو قرار می‌دهیم، این امکان وجود دارد که نتیجه‌ی جستجو همان کلمات کلیدی موجود در پرسش را شامل شود، ولی از لحاظ معنایی و زمینه‌ی کاربرد کاملاً با هدف فرد جستجوکننده فاصله داشته باشد، برای رفع این موضوع، از روش‌های مختلفی می‌توان استفاده کرد مثل Page Rank model و hub and authority model که براساس ارتباط میان اسناد است. ما بیشتر سعی داریم که این ارتباط را بر اساس ارتباط کلمات کلیدی موجود در پرسش، در سند مورد جستجو پیدا کنیم. [۵]

الگوریتم plane sweep بر روی لیستی از آفست‌های مرتب‌شده که مکان کلمه‌ی مورد جستجو را در سند نشان می‌دهند، اعمال می‌شود. در این الگوریتم ما نیاز داریم که اندازه بازه‌ی مورد جستجو مشخص باشد، برای این منظور یک محدوده‌ی بحرانی[۱] تعریف می‌شود. سایز محدوده، تعداد کلماتی است که در آن قرار دارد، پیداکردن محدوده‌های بحرانی منجر به پیداکردن محدوده‌های کاندید[۲] در سند مورد جستجو می‌شود به گونه‌ای که هیچ محدوده‌ی دیگری را شامل‌نشود.(محدوده‌ی کاندید، محدوده‌ای است که حداقل k کلمه‌ی کلیدی موجود در پرسش را شامل شود.)

ما در عملیات بازیابی اطلاعات معمولاً پرسش را می‌دانیم، بدین جهت تنها سعی داریم روابطی در داده‌ها با توجه به کلمه‌ی مورد جستجو پیدا کنیم، تا زمانی که حتی فضای جستجو بزرگ باشد، به پاسخ بهینه برسیم. از اینرو نیاز به روش‌هایی داریم که بتواند با کاهش تعداد مقایسه‌ها عمل جستجو را سریع‌تر انجام دهد. برای این منظور ما سعی داریم تا با پیداکردن کلمات تکراری پشت‌سرهم در لیستی از کلمات با آفست مشخص که خروجی مرحله‌ی پیش‌پردازش در الگوریتم plane sweep می‌باشد و شمارش آن‌ها و ذخیره‌سازی تعداد، و همچنین حذف کلیدواژه‌های غیرمفید در جستجو با استفاده از جستجوی بخشی، تعداد دفعات مقایسه را کاهش دهیم.

هدف این رهیافت ارائه روشی بر اساس الگوریتم plane sweep با در نظر گرفتن پارامترهای مؤثر

[۱] critical range
[۲] candidate range

مقدمه

می باشد، که علاوه بر ارائه پاسخ مناسب، سرعت یافتن پاسخ را افزایش دهد.

۱-۲- تعریف مسأله و بیان سوال‌های اصلی تحقیق

کاهش زمان مورد نیاز به منظور پاسخگویی به پرسش کاربر، از اهداف مهم هر موتور جستجو می‌باشد. با در نظرگرفتن ارتباط میان کلمات در الگوریتم plane sweep و کاهش تعداد مقایسه با استفاده از حذف تکرار و کلمات غیرمفید در جستجو، زمان پاسخگویی را کاهش دادیم به این ترتیب جستجو، تنها بر روی بازه‌های هدف انجام می‌شود و در نتیجه الگوریتم فوق دارای نتایج عملی کارآمدی می‌باشد.

سئوالات اصلی مورد نظر در بحث، در چهارچوب زیر می‌گنجد:

- ✓ آیا می‌توان یک الگوریتم بهینه برای جستجوی پرسش، بر اساس ارتباط مناسب میان پرسش و اسناد مورد جستجو ارائه نمود، به گونه‌ای که زمان اجرای الگوریتم کاهش یابد؟
- ✓ چگونه می‌توان با استفاده از داده‌های تکراری در سند زمان الگوریتم را بهبود بخشید؟
- ✓ چگونه می‌توان الگوریتم را به‌نحوی تنظیم نمود تا تعداد مقایسه‌های الگوریتم کاهش یابد؟
- ✓ چگونه می‌توان روند یافتن پاسخ (همگرایی پاسخ) توسط الگوریتم plane sweep را افزایش داد؟

۱-۳- فرضیه‌ها

- ➢ با توجه به راهکارهای ارائه شده و در نظرگرفتن خصوصیات جستجوی جمله در سند، قصد بر ارائه الگوریتمی، برای کاهش تعداد مقایسه‌ها در سند می‌باشد.

مقدمه

- یافتن راه‌حلی برای کاهش تعداد مقایسه‌ها در سند به قصد کاهش زمان اجرای الگوریتم.
- ارائه الگوریتم جستجو برای یافتن راه‌حل، که دارای پیچیدگی زمانی کمتری نسبت به الگوریتم‌های مورد مطالعه دیگر باشد.
- در نظر گرفتن کلمات تکراری پشت‌سرهم در سند.
- ایجاد ارتباط مناسبی میان پرسش و اسناد مورد جستجو.
- افزایش سرعت یافتن پاسخ (همگرایی پاسخ ها)، با استفاده از در نظرگرفتن کلمات کلیدی و عملیات موجود در الگوریتم plane sweep.
- درنظرگرفتن بازه‌های هدف در جستجو برای کاهش تعداد مقایسه.

۱-٤- اهداف تحقیق

به‌دلیل اهمیت مساله جستجوی جمله در سند به خصوص با در نظرگرفتن کلمات تکراری پشت‌سرهم و تاثیر آن در کارایی سیستم، برآنیم تا الگوریتمی ارائه دهیم که با در نظرگرفتن پارامترهای مهم، پرسش مورد نظر کاربر را در حجم کثیری از اسناد، جستجو کرده و با توجه به نیاز به پاسخ‌دهی سریع در محیط عملیاتی، راه‌حلی مناسب به منظور کاهش تعداد مقایسه‌ها یافت شود.

- زمان اجرای الگوریتم در حجم بالای داده کاهش یابد.
- پیچیدگی الگوریتم plane sweep نسبت به الگوریتم‌های مورد مطالعه موجود کاهش یابد.
- سرعت همگرایی پاسخ‌ها در الگوریتم plane sweep افزایش یابد.
- افزایش رضایت کاربران به علت افزایش سرعت جستجو و ارتباط نتایج با جمله‌ی مورد جستجو.

۱-۵- روش تحقیق

روش انجام تحقیق به صورت مروری از طریق مطالعه طرح‌ها و روش‌های مختلف موجود در راستای موضوع تحقیق و نیز به صورت تطبیقی از طریق مقایسه با روش‌های مرتبط موجود می‌باشد.

۱-٦- مراحل انجام تحقیق

- شناسایی و بیان کامل مساله (تدوین مجموعه‌ای توصیفی از ویژگی‌ها، شباهت‌ها و تفاوت-های میان الگوریتم‌های جستجو و قابلیت‌های آن‌ها).
- مطالعه پیرامون مساله داده‌های تکراری در سند و جستجوی تکراری همسایگی در بازه‌های غیرهمپوشان (تدوین مجموعه، چالش‌ها، فرصت‌ها، و پارامترهای مؤثر در جستجو، شناسایی و مقایسه نقاط قوت و ضعف، و طبقه‌بندی روش‌ها و الگوریتم‌های ارائه شده و استفاده از ویژگی‌های بهینه آن‌ها).
- ارائه روشی مبتنی بر الگوریتم plane sweep (ارائه الگوریتم IPNWPSR با استفاده از داده-های تکراری موجود در سند).
- ارزیابی الگوریتم پیشنهادی و مقایسه آن با الگوریتم‌های دیگر (نتایج ارزیابی الگوریتم‌ها با تست شرایط مختلف)
- نتایج و ارزیابی پروژه.