



دانشگاه سям نور
په

دانشکده فني و مهندسي
گروه علمي مهندسي کامپيوتر و فناوري اطلاعات
پايان نامه کارشناسي ارشد
رشته مهندسي کامپيوتر - گرايش نرم افزار

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 آغازیان

آذر ۱۳۹۱



دانشگاه سям نور
په

دانشکده فني و مهندسي
گروه علمي مهندسي کامپيوتر و فناوري اطلاعات
پايان نامه کارشناسي ارشد
رشته مهندسي کامپيوتر - گرايش نرم افزار

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 آغازیان

آذر ۱۳۹۱





جمهوری اسلامی ایران
وزارت علوم، تحقیقات و فناوری

مرکزی



صور تجلسه 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 در روز سه شنبه مورخ ۱۳۹۱/۰۹/۰۷ ساعت ۱۱ در مرکز ری برگزار شد و پس از بررسی پایان نامه مذکور با نمره به عدد ۵۰...۱۸... به حروف هجده و هشتم و با درجه عالی... مورد قبول واقع شد نشد

هیات داوران:

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

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.tpu.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.tpu.ac.ir

آیا پایان نامه مذکور نیاز به اصلاحات دارد؟

اینجانب دانشجوی ورودی سال مقطع کارشناسی ارشد 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 نام خانوادگی دانشجو
تاریخ و امضاء

اینجانب دانشجوی ورودی سال مقطع کارشناسی ارشد 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 نام خانوادگی 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، به دلیل رهنمودها و تشویق‌های موثرشان، که زمینه‌ساز این تحقیق بوده است، صمیمانه قدردانی می‌کنم.

در انتها از پدر و مادر عزیزم به خاطر 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 پشت سرهم 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: الگوریتم plane sweep ، داده‌های تکراری، بازیابی متن، جستجوی proximity، جستجوی بخشی.

فهرست جدول‌ها.....	ث
فهرست شکل‌ها.....	ج
فصل اول مقدمه.....	۱
۱-۱- پیشگفتار.....	۲
۲-۱- تعریف مسأله و بیان سوال‌های اصلی تحقیق.....	۳
۳-۱- فرضیه‌ها.....	۴
۴-۱- اهداف تحقیق.....	۴
۵-۱- روش تحقیق.....	۵
۶-۱- مراحل انجام تحقیق.....	۵
۷-۱- ساختار پایان‌نامه.....	۶
فصل دوم جمع‌آوری داده‌ها و الگوریتم مجاورت.....	۷
۱-۲- جمع‌آوری داده‌ها.....	۸
۱-۱-۲- فایل‌های بزرگ.....	۸
۲-۱-۲- مخزن.....	۹
۳-۱-۲- شاخص سند.....	۹
۴-۱-۲- واژه‌نامه.....	۱۰
۵-۱-۲- لیست‌های برخورد.....	۱۱
۶-۱-۲- شاخص‌های روبه‌جلو.....	۱۲
۷-۱-۲- شاخص معکوس.....	۱۳
۱-۷-۱-۲- نحوه‌ی ایجاد شاخص معکوس.....	۱۴
۲-۷-۱-۲- انواع شاخص معکوس.....	۱۷
۱-۲-۷-۱-۲- شاخص معکوس در سطح رکورد.....	۱۷
۲-۲-۷-۱-۲- شاخص معکوس در سطح کلمه.....	۱۸
۲-۲- الگوریتم‌های جستجو.....	۲۰
۱-۲-۲- الگوریتم Page Rank.....	۲۱
۲-۲-۲- الگوریتم HITS.....	۲۴
۳-۲-۲- متن لنگری.....	۲۶

۲۷ ۴-۲-۲- جستجوی مجاورتی
۲۸ ۱-۴-۲-۲- اهمیت جستجوی مجاورتی
۲۸ ۲-۴-۲-۲- الگوریتم جستجوی مجاورت k -کلمه
۳۱ الگوریتم plane sweep
۳۲ ۱-۳- مقدمه
۳۲ ۲-۳- تعریف مسئله
۳۴ ۳-۳- الگوریتم Plane Sweep
۴۰ ۴-۳- نتیجه‌گیری
۴۱ الگوریتم جدید IPNWPSR
۴۲ ۱-۴- الگوریتم جدید مبتنی بر الگوریتم plane sweep
۴۲ ۱-۱-۴- مقدمه‌ای بر الگوریتم WPSR
۴۳ ۲-۱-۴- تعریف الگوریتم WPSR
۴۹ ۳-۱-۴- مثالی از الگوریتم WPSR
۵۰ ۲-۴- الگوریتم IPNWPSR
۵۸ ۱-۱-۴- مراحل اجرای الگوریتم IPNWPSR
۶۳ ۲-۱-۴- مثال‌هایی از الگوریتم IPNWPSR
۶۷ ۳-۴- نتیجه‌گیری
۶۸ فصل پنجم شبیه‌سازی و ارزیابی
۶۹ ۱-۳- مقدمه
۶۹ ۲-۳- شبیه‌سازی
۷۵ ۳-۳- جمع بندی

۷۶.....	فصل ششم نتیجه‌گیری و کارهای آینده.....
۷۷.....	۱-۶- مقدمه.....
۷۷.....	۲-۶- یافته‌های تحقیق.....
۷۸.....	۳-۶- نوآوری تحقیق.....
۷۹.....	۴-۶- پیشنهادهایی برای تحقیقات آینده.....
۸۰.....	فهرست مراجع و منابع.....
۸۴.....	واژه‌نامه انگلیسی به فارسی.....
۸۷.....	واژه‌نامه فارسی به انگلیسی.....

فهرست جدول‌ها

صفحه	عنوان
۶۹	جدول ۵-۱ نتایج آزمایش بر روی آفست‌ها در حالت ۳-کلیدواژه و $W_{sim} = 0.4$
۷۱	جدول ۵-۲ نتایج آزمایش بر روی آفست‌ها.....
۷۲	جدول ۵-۳ نتایج آزمایش بر روی آفست‌ها در حالت ۴-کلیدواژه‌ای و $R_D = 0.5$ و $W_{sim} = 0.5$
۷۳	جدول ۵-۴ نتایج آزمایش بر روی آفست‌ها در حالت ۳-کلیدواژه‌ای و $R_D = 0.6$ و $W_{sim} = 0.6$
۷۴	جدول ۵-۵ نتایج آزمایش بر روی آفست‌ها با اندازه و ضریب تشابه متفاوت.....

فهرست شکل‌ها

صفحه	عنوان
۱۴	شکل ۲-۱ عملیات شاخص‌گذاری سستی
۱۵	شکل ۲-۲ پیاده‌سازی فایل معکوس با استفاده از آرایه مرتب شده
۱۶	شکل ۲-۳ نمایش ایجاد آرایه مرتب‌شده‌ی فایل معکوس
۱۸	شکل ۲-۴ اندیس سطح سند [۶]
۱۹	شکل ۲-۵ اندیس معکوس سطح کلمه [۶]
۲۹	شکل ۲-۶ تعیین بازه‌های مینیمال و غیرمینیمال در سند DDDCABCD
۳۰	شکل ۲-۷ الگوریتم جستجوی مجاورتی [۶]
۳۴	شکل ۳-۱ یک مثال ساده [۹]
۳۵	شکل ۳-۲ مثالی از لیست ادغامی [۹]
۳۷	شکل ۳-۳ مثالی از بازه بحرانی و کاندید [۹]
۳۹	شکل ۳-۴ مثالی از لیست ادغام شده در الگوریتم اصلاح شده Plane Sweep
۴۰	شکل ۳-۵ یافتن راه‌حل برای مثال ۱
۴۸	شکل ۴-۱ فلوچارت الگوریتم WPSR
۴۹	شکل ۴-۲ مثالی از لیست ادغام شده در الگوریتم اصلاح شده Plane Sweep
۵۰	شکل ۴-۳ یافتن راه‌حل برای مثال ۱
۵۴	شکل ۴-۴ فاکتور فاصله میان دو کلیدواژه‌ی مینیمم
۵۶	شکل ۴-۵ جستجو بر روی سند { ABCCCABCCBACBBBCBA }
۵۷	شکل ۴-۶ مثالی از جستجو با در نظر گرفتن مینیمم تکرار کلیدواژه
۵۸	شکل ۴-۷ جستجوی بخشی بر روی لیست آفست
۵۹	شکل ۴-۸ الگوی عمومی انتخاب پاسخ در الگوریتم IPNWPSR
۶۰	شکل ۴-۹ مراحل جستجوی الگوریتم IPNWPSR
۶۲	شکل ۴-۱۰ فلوچارت الگوریتم IPNWPSR

- شکل ۴-۱۱ جستجو بر روی بازه‌ی $K_1 \in \{0,6,8,10\}$ ----- ۶۳
- شکل ۴-۱۲ جستجو بر روی بازه‌ی $K_2 \in \{1,11\}$ ----- ۶۳
- شکل ۴-۱۳ جستجو بر روی بازه‌ی $K_3 \in \{5,7,9\}$ ----- ۶۴
- شکل ۴-۱۴ نتیجه جستجوی الگوریتم IPNWPSR بر روی بازه‌ی I_1, I_2, I_3 ----- ۶۴
- شکل ۴-۱۵ نمونه ایده‌آل از الگوریتم IPNWPSR ----- ۶۵
- شکل ۴-۱۶ نمونه‌ای از الگوریتم IPNWPSR که مشابه Plane sweep عمل می‌کند ----- ۶۵
- شکل ۴-۱۷ جستجو بر روی بازه‌های K_1, K_2, K_3 ----- ۶۶
- شکل ۴-۱۸ نتیجه جستجوی IPWPSR در بازه‌های ----- ۶۷
- شکل ۵-۱ نتایج مقایسه میان الگوریتم WPSR و plane sweep بر روی آفست‌ها در حالت ۳-کلیدواژه و $W_{sim} = 0.4$... ۷۰
- شکل ۵-۲ تعداد مقایسه‌ها در الگوریتم‌های WPSR و IPNWPSR با پرسش ۳-کلیدواژه و $W_{sim} = 0.4$ ----- ۷۱
- شکل ۵-۳ تعداد مقایسه‌ها در الگوریتم‌های plane sweep و IPNWPSR در حالت ۴-کلیدواژه‌ای ----- ۷۲
- شکل ۵-۴ تعداد مقایسه‌ها در الگوریتم‌های plane sweep و IPNWPSR در حالت ۳-کلیدواژه‌ای ----- ۷۳
- شکل ۵-۵ نمودار آزمایش بر روی آفست‌ها با اندازه و ضریب تشابه متفاوت ----- ۷۵

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

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 شامل نشود. (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

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 نظر گرفتن 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 به نحوی 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 نظر گرفتن خصوصیات جستجوی جمله در سند، قصد بر ارائه الگوریتمی، برای کاهش تعداد مقایسه‌ها در سند می‌باشد.

- 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 پشت سرهم در سند.
- ایجاد ارتباط مناسبی میان پرسش و اسناد مورد جستجو.
- افزایش سرعت یافتن پاسخ (همگرایی پاسخ‌ها)، با استفاده از در نظر گرفتن کلمات کلیدی و عملیات موجود در الگوریتم plane sweep.
- در نظر گرفتن بازه‌های هدف در جستجو برای کاهش تعداد مقایسه.

۱-۴- اهداف تحقیق

- به دلیل اهمیت مساله جستجوی جمله در سند به خصوص با در نظر گرفتن کلمات تکراری پشت سرهم و تاثیر آن در کارایی سیستم، برآنیم تا الگوریتمی ارائه دهیم که با در نظر گرفتن 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.

۱-۶- مراحل انجام تحقیق

- شناسایی و بیان کامل مساله (تدوین مجموعه‌ای توصیفی از ویژگی‌ها، شباهت‌ها و تفاوت‌های 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 داده‌های تکراری موجود در سند).
- ارزیابی الگوریتم پیشنهادی و مقایسه آن با الگوریتم‌های دیگر (نتایج ارزیابی الگوریتم‌ها با تست شرایط مختلف)
- نتایج و ارزیابی پروژه.