به نام خدا

مستندات فاز سوم پروژه درس بازیابی پیشرفته اطلاعات

تهیه کنندگان:

روح اله جهنده

محمد رضا خرمی

1 - بیتشرین page rank مربوط به Wikipedia:About است

2 - نتیجه ی کویری information در صورتی که از page rank استفاده نکنیم تفاوت بسیار زیادی با حالت قبل آن دارد

3 - قابلیت ها و تکنولوژی های استفاده شده:

a. Crawl کردن تعداد دلخواهی صفحه با استفاده از JSOUP
b. Crawlکردن سایت مشخص و یا آدرس هایی مشخص با ارائه یک regular expression
c. ذخیره کردن مشخصات صفحات و پیوندهای میان صفحات در پایگاه داده mysql با استفاده از hibernate(که بعد ها میتوان راحت تر قابلیت های دیگری را به این موتور جستجو استفاده کرد مثلا میتوان تیتر صفحات را در پایگاه داده ذخیره کرد و در فرآیند جستجو و ایندکس گذاری به آن ضریب داد. )
d. استفاده از کتابخانه ی jung برای بدست آوردن page rank ها که در نتیجه می توان با استفاده از پارامتر alpha الگوریتم page rank را smooth کنیم. همچنین با توجه به اینکه گراف جهت دار لینکهای صفحات به هم ساخته شده است می توان با استفاده از این کتابخانه آن را نمایش گرافیکی داد.
e. استفاده از lucene برای ایندکس گذاری و جستجو.

4 - جهت تصحیح باید موارد زیر انجام شود:

a. نصب بودن mysql و ایجاد پایگاه داده ای به نام mir_p_ph3 و دیگر مشخصات پایگاه داده که در فایل hibernate.cfg.xml است.
b. تنظیم کردن آدرس مطلق مربوط به فایل های پیکره و ایندکس در کلاس SimpleFileIndexer
c. امکان crawl با استفاده از UI به دلیل احتمال حرکت سهوی اشتباه در کد مربوط searchBox.java لغو شده است باید یک false را به true تغییر دهید.
d. فایل شروع ابتدایی SearchBox.java است.

5 - مشکلات و نقاط ضعف:

a. Page Rank بسیار بالای بعضی صفحات مثل about که نتیجه ی یک کویری را نامطلوب میکند. یکی از راه های موثر فکر می کنم زیاد کردن صفحات crawl شده است.
b. مشکلاتی که در درس به آن اشاره شد مانند لینک دهی های عمدی یک گروه به هم با هدف اسپم و یا تکرار عمدی واژه ها کاملا می‌تواند در عملکرد موتور جستجوی نوشته شده کاملا اخلال ایجاد کند.
c. صفحات duplicate تشخیص داده نشده اند و کاربر را با نتایج مشابه ناراضی می‌کند
d. استفاده از ajax (تولید محتواها توسط جاواسکریپت) در سایت ها کاملا می تواند در این موتور جستجو اخلال ایجاد کند برای مثال سایت plus.google.com
e. با توجه به اینکه زمان بندی در این موتور جستجو رعایت نشده ممکن از بعضی از سایت ها آن را حمله کننده تشخیص دهند و آن را block کنند که این اتفاق عملکرد crawler را با مشکل جدی مواجه می‌کند
f. در این موتور جستجو فایلهای robot در نظر گرفته نمی شوند و ممکن است با نقض قوانین سایت، مشکلی شبیه مورد قبل اتفاق بیافتد.

نحوه ی عملکرد به صورت خلاصه:

ابتدا یک صفحه دانلود میشود سپس لینک هایش به صفحات دیگر جدا میشود (البته تا جایی به ماکزیمم تعداد صفحه نرسیم) و در صف قرار می گیرند(به ازای هر لینک جدید یک ردیف در پایگاه داده اضافه می شود و همچنین در جدول دیگری ارتباطات جهت دار ذخیره می شود) و متنش در یک فایل ذخیره میشود و این کار تکرار می شود تا به تعداد مشخص شده برسد.

سپس گراف جهت دار لینکها بدست آورده می شود و pagerank صفحات بدست می آید.

سپس فایلها ایندکس گذاری می شوند.

سپس کویری داده شده و نتایج گرفته می شود.