پروژه پایانی درس بازیابی پیشرفته اطلاعات

رتبه‌بندی وبلاگ‌های فارسی

(۸۴/۱۰/۹)

۱. رتبه‌صفحات در وب

وب، ساختار داده‌ای ویژه‌ای‌ست و خصوصیات منحصر به فردی دارد. یکی از ویژگی‌های بارز آن عدم وجود کنترل مرکزی بر فرآیند تولید و نشر است. به همین دلیل صفحات متعددی در وب وجود دارند که اطلاعات موجود در آنها ناقص، نادرست و جعلی می‌باشد. از طرفی به علت آنکه در مساله جستجو میزان اعتبار نتایج اهمیت بالایی دارد، پژوهش‌های گوناگونی در حوزه اعتبار سنجی صفحات موجود در وب انجام شده‌است. یکی از مهمترین و موفق‌ترین نتایج این تحقیقات، الگوریتم PageRank است که توسط بنیانگذاران موتور جستجوی گوگل ارائه گردیده است و با کمک آن می‌توان بین صفحات وب ترتیب برقرار نمود. بر طبق این الگوریتم هرچه تعداد صفحاتی که به یک صفحه خاص لینک داده‌اند بیشتر باشد، آن صفحه رتبه بالاتری خواهد داشت. در ضمن هر چه تعداد لینک-های خروجی از یک صفحه بیشتر باشد، ارزش این لینک‌ها کمتر خواهد بود.

۲. شرح پروژه

هدف از این پروژه، محاسبه PageRank برای وبلاگ‌های فارسی است. برای سادگی، فقط وبلاگ‌های میزبانی شده در پرشین‌بلاگ مدنظر است. (الگوی کلی آدرس آنها به فرم http://weblogName.persianblog.com می‌باشد که در آن weblogName یک رشته از کاراکتراهای مجاز است). این پروژه را می‌توان به بخش‌های زیر تقسیم نمود:

۲،۱. *کشف وبلاگ‌ها:* در این قسمت باید یک خزشگر (Crawler) پیاده شود تا عملیات کشف وبلاگ‌ها را انجام دهد. این خزشگر با شروع از یک وبلاگ خاص، آدرس آنرا ذخیره نموده و پس از استخراج لینک‌های خروجی به صورت عمق‌اول یا سطح‌اول این کار را تکرار می‌کند. خروجی این عملیات را برای استفاده در بخش‌های بعد، در یک پایگاه‌داده ذخیره نمایید. با توجه به فرمت کلی آدرس این وبلاگ‌ها بهتر است در هنگام ذخیره‌سازی و کار با آنها به جای کل URL فقط از weblogName ها استفاده کنید.

۲،۱،۱. تعداد وبلاگ‌های موجود در پرشین‌بلاگ بین یک‌صد هزار تا پانصدهزار برآورد می‌شود. نمره کامل این بخش به گروهی تعلق می‌گیرد که بیشترین تعداد وبلاگ را کشف کرده باشد.

۲،۱،۲. با توجه به این که گراف‌های وبی، گراف‌های قویا همبندی نیستند، اکیدا توصیه می‌شود که از چند نقطه‌ی شروع متفاوت استفاده کنید. یک راه‌حل مناسب، انتخاب تعدادی از لینک‌های موجود در طبقات مختلف "فهرست کاربران" در صفحه اصلی پرشین‌بلاگ (http://www.persianblog.com) است.

۲،۱،۳. به علت آنکه در این پروژه تاکید فقط بر گراف پرشین‌بلاگ است، در صورتی که وبلاگ‌های موجود در آن به یک صفحه خارج از پرشین‌بلاگ لینک داده باشند، این لینک‌های خروجی نباید پردازش شوند (هرس شاخه). اما لیست این صفحات خارجی را همراه با تعداد لینک‌های داده شده به آن (از وبلاگ‌های پرشین‌بلاگ) در یک جدول جداگانه نگهداری نمایید.

۲،۱،۴. یکی از نکات مهمی که باید مورد توجه قرار بگیرد، اسامی مختلف یک وبلاگ معین است. به عنوان مثال ممکن است به یک وبلاگ خاص به دو صورت متفاوت http://www.weblogName.persianblog.com و http://weblogName.persianblog.com لینک داده‌شده‌باشد. مثال دیگر در این زمینه امکان وجود لینک از یک وبلاگ به یک یادداشت خاص در یک وبلاگ دیگر است (مثلا http://weblogname.persianblog.com/1384_10_weblogname_archive.html#447158 و یا http://www.weblogname.persianblog.com/#4450955) از پردازش این دسته‌از لینک‌ها صرفنظر کنید.

۲،۲. *محاسبه رتبه وبلاگ‌ها:* پس از آماده‌سازی پایگاه‌داده مربوطه، مرحله بعد محاسبه رتبه صفحات می‌باشد. مقدار مجاز برای PageRank در این پروژه بین ۰ تا ۱۰ در نظرگرفته می‌شود. در ابتدا برای همه صفحات این مقدار را برابر ۱۰ قرار داده و سپس مطابق الگوریتم گفته‌شده، در یک فرآیند تکرار شونده (حداقل ۱۰تکرار) مقدار PageRank متناظر با هر صفحه را به‌روز نمایید. در صورتی که مقادیر PageRank همگرا نشود تعداد این تکرارها را افزایش دهید.

۲،۲،۱. از ساده‌ترین و پایه‌ای‌ترین حالت الگوریتم PageRank برای این محاسبات استفاده کنید.

$$R(A) = \sum_{(B,A)\in G} R(B) / outdegree(B)$$

۲،۳. *نمایش نتایج:* در این قسمت باید یک واسط کاربر مبتنی بر وب پیاده شود. یکی از امکانات ضروری این واسط، نمایش مقدار رتبه وبلاگ و لیست همه لینک‌های ورودی وخروجی آن پس از دریافت نام وبلاگ – همان weblogName – است. همچنین این واسط باید به کاربر اجازه دهد تا با وارد کردن یک (دو) عدد به عنوان یک رتبه (یک بازه از رتبه‌ها)، وبلاگ(های) متناظر را پیدا کند. لازم به ذکر است که در نهایت مقدار PageRank صفحات، یک عدد گویا بین ۰ تا ۱۰ می‌باشد اما رتبه صفحات یک عدد طبیعی بین ۱ تا تعداد وبلاگ‌های کشف شده‌است.

۳. نکات مهم

- نمره این پروژه ٪۲۰ از نمره پایانی درس را تشکیل می‌دهد.
- نحوه انجام پروژه به صورت گروهی می‌باشد (۴ نفری). نمایندگان هر گروه اسامی اعضای گروه خود را تا روز امتحان نهایی به حل‌تمرین درس اعلام کنند.
- موعد نهایی تحویل جدول نتایج **۸۴/۱۱/۱۰** می‌باشد. با توجه به محدودیت زمانی ارسال نمرات به آموزش این مهلت قابل تمدید نیست.
- قبل از روز تحویل هر گروه باید یک گزارش از پروژه تحویل دهد. مواردی که باید در این گزارش ذکر شود:
  - نحوه تقسیم کار بین اعضای گروه
  - مشکلات فنی پیش‌آمده در حین اجرا و نحوه حل آنها
  - صد وبلاگ اول
  - صد وبلاگ آخر
  - صد صفحه پرطرفدار خارج از پرشین‌بلاگ (بر حسب شمارش لینک‌های داده شده به آن از درون وبلاگ‌ها)

- در تعریف این پروژه ممکن است تغییراتی محدودی اعمال شود. آخرین نسخه را می‌توانید از وب‌سایت درس دریافت نمایید.
- برای پرسیدن سوالات خود از بخش Discussion Area در وب سایت درس استفاده کنید.

توضیحات جدید

(۸۴/۱۰/۳۰)

- در هنگام ساخت گراف وب، از منظور کردن لینک‌های مرده (که به وبلاگ‌های پاک‌شده اشاره می‌کنند) خودداری نمایید.
- در گزارش نهایی موارد زیر را نیز لحاظ نمایید:
  - تعداد نهایی وبلاگ‌های یافت شده
  - اسامی صد وبلاگ اول از لحاظ تعداد لینک‌های ورودی
  - لیست نقاط شروع در عملیات خزش و توضیح در مورد نحوه انتخاب آنها

بخش نمره اضافه

(۸۴/۱۱/۳)

- برای کسب نمره اضافی در این پروژه می‌توانید الگوریتم HITS را روی گراف پرشین‌بلاگ پیاده نمایید. سیستم باید امکان وارد نمودن درخواست کاربر را در قالب یک کلمه، یک جمله و یا یک عبارت فراهم نماید، سپس با کمک موتور جستجوی گوگل ۵۰ وبلاگ (از مجموعه پرشین‌بلاگ) که به این پرسش مرتبط‌تر باشند را انتخاب نموده و به عنوان مجموعه ریشه (Root Set) در نظر می‌گیرد. در مرحله بعد باید با استفاده از گراف ساخته‌شده در پروژه اصلی، مجموعه ریشه را توسعه داده تا مجموعه پایه (Base Set) حاصل شود. قسمت آخر کار، محاسبه بردارهای Hub و Authority برای این مجموعه‌است.
- بارم بخش‌اضافه، ۱۰٪ از نمره نهایی درس خواهد بود. (۲ نمره)
- در مورد بخش نمره اضافه نکات زیر را مدنظر قرار دهید:
  - توصیه می‌شود برای جمع‌آوری نتایج اولیه از امکان جستجوی درون-سایتی گوگل استفاده کنید (مثلا پرسش کاربر به فرمت persianblog.com: SampleQuery تبدیل شده و به گوگل ارسال شود).
  - توجه نمایید که مجموعه‌های ریشه و پایه فقط و فقط باید روی صفحات پرشین‌بلاگ ساخته‌شود.
  - در هنگام محاسبه بردارهای H و A، الگوریتم HITS را تا ۱۰ قدم تکرار کنید.
  - واسط کاربری باید دارای یک بخش ورودی برای دریافت پرسش‌کاربر باشد. در بخش خروجی، برای هر پرسش باید لیست ۵۰ وبلاگ ریشه، بهترین Hubها و بهترین Authorityها (هرکدام ۱۰ مورد) به کاربر نشان داده شود.
  - گزارش کتبی بخش‌نمره‌اضافه باید به صورت مجزا از گزارش اصلی تهیه و تحویل داده شود
  - در این گزارش باید برای هر کدام از پرسش‌های زیر نتایج گفته شده آورده شود
    - جنگ
    - *"روزنامه شرق"*
    - *جشنواره بین‌المللی فیلم فجر*

توضیحات جدید

(۸۴/۱۱/۸)

- استفاده از پایگاه‌داده برای ذخیره‌سازی گراف الزامی می‌باشد. جهت سهولت کار و جلوگیری از ایجاد ازدحام روی سرور دانشکده، توصیه می‌شود که ابتدا اطلاعات را درون فایل ریخته و سپس به صورت offline وارد پایگاه‌داده نمایید. همچنین طراحی امکان pause در تسهیلِ فرآیند خزش بسیار موثر می‌باشد.
- موارد اضافه‌شده به گزارش کتبی:
  - مستندات کلی طراحی پایگاه‌داده (شامل جداول، فیلدها).
  - تعداد زیرگراف‌های مجزا در مجموعه پرشین‌بلاگ (بخش‌های ناهمبند) همراه با ذکر نقاط شروع هرکدام.
- در مورد بخش‌نمره‌اضافه:
  - در صورتی بروز هرگونه مشکلی در کار با گوگل، می‌توانید از یاهو استفاده کنید.
  - بردارهای Hub و Authority را برای کل گرافِ پرشین‌بلاگ محاسبه کنید (فرض کنید که Base Set کل گراف است) و لیست صد Hub اول و صد Authority اول را در گزارش خود ذکر کنید.
- موعد تحویل حضوری این پروژه، روز ۸۴/۱۱/۱۵ می‌باشد. درمورد تحویل این پروژه به نکات زیر توجه کنید:
  - مسولان هرکدام از تیم‌ها موظفند که مستندات پروژه خود را حداکثر تا روز ۸۵/۱۱/۱۴ به حل تمرین درس ایمیل بزنند.
  - در روز تحویل هر تیم باید یک CD ، حاویِ برنامه‌ها، مستندات و فایل‌های ذخیره شده (Dump) پایگاه‌داده ارائه نماید.
  - بعد از دریافت مستندات پروژه‌ها، برنامه زمان‌بندی تحویل آنها متعاقبا اعلام خواهد شد.
  - نمرات این درس روز ۱۶ بهمن‌ماه به آموزش ارسال خواهد شد، فلذا امکان تمدید زمان تحویل به‌هیچ وجه وجود ندارد.