مبانی بازیابی اطلاعات و جستجو

سید مهدی وحیدی پور

ارایه اول: معرفی درس

مبانی بازیابی اطلاعات و جستجو

- اطلاعات درس
  - زمان کلاس: یکشنبه ساعت ۱۲:۳۰ تا ۱۴ -- دوشنبه ساعت ۱۰ تا ۱۲
  - منابع درسی:
    - اسلایدهای درسی
    - مطالب گفته شده در کلاس
    - کتابهای مرجع درسی
  - نمره
    - حضور، تمرین، کوییز، پروژه ۴۰٪
    - میانترم ۳۰٪
    - پایانترم ۳۰٪
    - تبصره: دانشجویانی که بیش از ۴ جلسه غیبت نمایند، باید درس را حذف کنند.
  - وبگاه درس
- https://faculty.kashanu.ac.ir/vahidipour/fa/page/بازیابی و جستجو

اجزای درس بازیابی اطلاعات و جستجو

قسمت اول درس

درس ذخیره و بازیابی اطلاعات (کارشناسی در دوره قبل)

- مفاهیم فیلد، رکورد و فایل
- جدول، کلید اولیه و پایگاه داده‌ها
- ساختارهای اندیسی
- ...

قسمت دوم درس

درس بازیابی پیشرفته اطلاعات (کارشناسی ارشد)

- سند
- جمع‌آوری، ذخیره‌سازی، اندیس سازی
- کوئری و نتایج جستجو
- ...

مرور سریع درس

- درس ذخیره و بازیابی اطلاعات پیش نیاز درس پایگاه داده‌ها بوده است.
- درس بازیابی پیشرفته اطلاعات در دوره کارشناسی ارشد مطرح است.
- انواع جستجو:
  - جستجوی ترتیبی و مستقیم
  - جستجوی ctrl+f در یک فایل
    - روشهای string matching: سعی در افزایش سرعت جستجوی ترتیبی در متن
    - افزایش کارایی جستجو با کاراکترهای *، ?، + و... در عبارتهای جستجو
    - نتایج جستجو رتبه بندی Ranking ندارد
    - روشهای مبتنی بر شاخص مانند suffix tree: افزایش سرعت جستجو.
  - جستجو در پایگاه داده‌ها
    - کلید اولیه، فایل شاخص...
  - جستجو در وب (اسناد متعدد)
    - موتورهای جستجو
    - اسناد ساخت یافته نیستند
    - نتایج جستجو رتبه بندی می‌شوند.

مراجع درس

- م. ج. فولک و سایر همکاران، ساختار فایل‌ها، ۱۳۸۳.
- W. Bruce Croft, Donald Metzler, Trevor Strohman , *Search Engines: Information Retrieval in Practice,* Pearson Education, 2010.
- Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition), ACM Press Books, 2010.
- C. Manning, P. Raghavan, and H. Schütze , Introduction to Information Retrieval, Cambridge University Press, 2008.

فیلد

واحد اصلی داده ها فیلد است.

چهارروش متداول جدا کردن یک فیلد از فیلد بعدی:

- **شروع کردن فیلدها درطولهای قابل پیش بینی.**
- **شروع کردن هرفیلدی با نشانگرطول فیلد.**
- **قراردادن یک فاصل (delmitier) درانتهای هرفیلد برای جدا کردن آن ازفیلد بعدی.**
- **استفاده ازیک عبارت کلیدی برای شناسایی هرفیلد ومحتویات آن.**

رکورد

رکورد مجموعه ای از فیلدهاست.

بعضی از روش های سازماندهی رکوردهای فایل عبارتند از:

- **قابل پیش بینی کردن طول رکوردها بر حسب بایت.**
- **قابل پیش بینی کردن طول رکوردها بر حسب فیلدها.**
- **شروع هر رکورد با نشانگر طول، که تعداد بایتهای رکورد را نشان میدهد.**
- **استفاده از فایل دیگری برای آدرس شروع هر رکورد.**
- **قرار دادن فاصل در انتهای هر رکورد، برای جدا کردن آن از رکورد بعدی.**

- **رکورد با طول ثابت و متغیر**

رکورد

- کلید رکورد
  - برای جستجو بین رکورد ها باید یک شکل استاندارد برای کلیدها تعریف کنیم .
  - این شکل استاندارد را شکل کانونیک کلید می نامند.
  - این شکل بایستی منحصر بفرد باشد

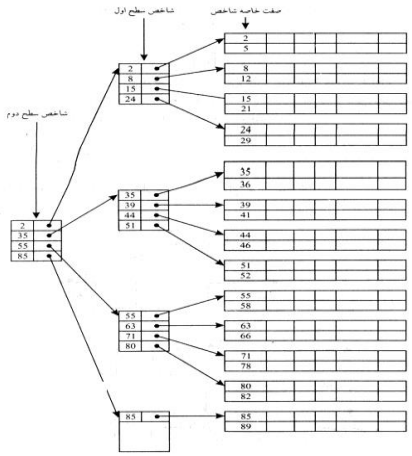
مدیریت فایل‌هایی از رکوردها

- دستیابی مستقیم
  - هنگامی که بتوانیم مستقیما به ابتدای یک رکورد برویم و آن را به حافظه وارد کنیم به آن رکورد دستیابی مستقیم داریم.
  - O(1)
  - توابع درهم‌ساز
- جستجوی ترتیبی
  - در این روش فایل رکورد به رکورد خوانده می شود تا رکوردی با یک کلید خاص پیدا شود.
  - O(N)
  - بلوک بندی می‌تواند کارایی جستجوی ترتیبی را افزایش دهد.
  - خوب است وقتی که
    - فایلهای اسکی که در آنها بدنبال یک الگو هستیم
    - فایلهایی با تعداد محدود از رکوردها
    - فایلهایی که بندرت نیاز به جستجو دارند

شاخص

- ترکیبی از دستیابی مستقیم و جستجوی ترتیبی
- منظور از شاخص مجموعه ای از عناصر شاخص است که به صورت جفت های(x , a) از داده هایی با طول ثابت است که به طور فیزیکی کنار هم قرار دارند . x نشانگر کلید و a نشانگر اطلاعات همراه با کلید است .
- فرض می کنیم خود شاخص انقدر بزرگ است که تنها بخش کوچکی از آن را می‌توان در یک لحظه در حافظه اصلی نگه داشت . بنا براین شاخص باید در یک حافظه جانبی ذخیره شود.
- فایل داده (رکوردهایی از فیلدها) + فایل شاخص
- بدون دستکاری محتویات فایل داده ,به فایل نظم و ترتیب می‌بخشند.

شاخص چند سطحی



- ساده‌ترین شاخص چند سطحی: درخت دودویی
- برای جستجوی دودویی لازم است تا رکوردهای داخل فایل داده بر اساس کلید اولیه مرتب شوند.
- در فایل اندیس کلیدهای اولیه در ساختار چند سطحی، در قالب درخت دودویی مرتب شده و از آن برای جستجو استفاده می‌شود.
- شاخص باعث می شود تا رکورد ها را به وسیله کلید آنها با سرعت زیادی پیدا کنیم .سرعت این کار در مقایسه با حالتی که جستجوی دودویی در یک فایل مرتب موجود در حافظه انجام می شود بیشتر است.