اهمیت موتورهای جستجو

- رتبه شماره یک موتورهای جستجو در اکثر کشورها را از نظر تعداد بازدید کننده
- بزرگترین شرکتهای حوزهٔ فناوری اطلاعات و دارای رتبهٔ اول درآمد در بین سایتهای اینترنتی در کشورهای مختلف
- ٪۸۰ کاربران اینترنت از موتور های جستجو برای دسترسی به اطلاعات استفاده میکنند

اهمیت موتورهای جستجو

نام	کشور	درآمد (میلیارد دلار)	درصد تبلـغات در درآمد	تعداد نیرو	سهم پرس و جو	تعداد پرس و جو در روز
گوگل	آمریکا	۶۶میلیارد	۹۰	۵۴۰۰۰	٪۷۰ در آمریکا	۱۰ میلیارد
بایدو	چین	۸ میلیارد	۶۰	۴۰۵۰۰	٪۸۶ در چین	۳ میلیارد
یاندکس	روسیه	۱میلیارد	اکثر در آمد	۵۴۰۰	٪۶۵ در روسیه	۲۰۰ میلیون
سزنم	جمهوری چک	۰/۱۳ میلیارد	اکثر درآمد	۱۱۰۰	٪۵۰ در جمهوری چک	۸ میلیون
نیور	کره جنوبی	حدود از ۲/۵ میلیارد	اکثر درآمد	۲۵۰۰	٪۷۳ در کره جنوبی	۴۰۰ میلیون

مزیتهای موتورهای جستجوی بومی

- شناخت بهتر زبان و فرهنگ
  - دلیل موفقیت یاندکس: فهم و پردازش بهتر زبان روسی
  - دسترسی کاربران بایدو به اطلاعات شبکهٔ مجاز چین
- ارزش افزودهٔ اطلاعاتی
  - سرویسهای محلی و بومی
- اجتماعی و فرهنگی
  - هدایت و راهبری کاربران به اطلاعات معتبر و مجاز
- دولت الکترونیک و اینترانت داخلی
  - نمونه موفق نیور در کره جنوبی
- بستری برای ساماندهی فضای مجازی
  - پاسخ به کاربران متناسب با نیاز اقشار (دانش آموزان، دانشجویان، صاحبان وب سایت)
  - مثال کره جنوبی (سرویسهای اختصاصی برای قشرهای مختلف)

خصوصیات پارسی‌جو

- **پوشش یک و نیم میلیارد سند فارسی با محتوای امن (درگاه رسمی)**
- **دارای خزشگری با قدرت خزش بیش از ده میلیارد سند به صورت متوالی**
  - **هوشمند در تشخیص اسناد مهم**
  - **خزش دوره ای در بازه های منظم**
- **نمایه سازی و پردازش سریع اطلاعات**
- **طراحیِ مبتنی بر بستر توزیع شده و مقیاس پذیر**
- **استفاده از پردازشگر هوشمند زبان فارسی**
  - **طراحی و پیاده سازی یک خطایاب هوشمند**
  - پیاده سازی پردازشگر متون فارسی
- **استفاده از روش رتبه بندی کارا**
  - **بهینه سازی و ارتقاء مداوم الگوریتم**

سرویس ها

سرویس	وضعیت کنونی
وب	• سرویس وب، در بین صفحات فارسی موجود در فضای مجازی جستجو می کند. با توجه به اینکه صفحات وب از طریق گذرگاه رسمی شبکه کشور دریافت می شوند، دارای محتویات پاک و پالوده اند. مزیت نسبی سرویس فوق نسبت به سرویس وب موتورهای دیگر، ارائه سرویس های دیگر نظیر سرویس عکس، آوا و ... در نتایج آن است.
تصویر	• این سرویس در بین تصاویر صفحات وب فارسی جستجو می کند و همانند سرویس وب دارای محتویات امن و مناسبی است. سرویس تصویر، خدماتی همانند پیدا کردن تصاویر مشابه، جستجو بر حسب اندازه و رنگ و صورت ارائه می دهد.
ویدئو	• سرویس ویدئو امکان جستجو و پخش فایل های تصویری موجود در صفحات وب فارسی را به صورت مستقیم، در اختیار کاربران قرار می دهد.
خبر	• این سرویس شامل بخش سرویس خبر وب و اپلیکیشن موبایلی خبرجو می باشد. • این سرویس امکان خزش در بیش از شصت خبرگزاری برتر کشور به صورت هوشمند و خودکار، جستجو در اخبار، تصاویر و فیلمهای خبری و دسته بندی آن ها را برای کاربر میسر می سازد.
آوا	• سرویس آوا امکان جستجوی صوت و موسیقی را به صورت مستقیم از فایلهای صوتی موجود در سطح وب فارسی در اختیار کاربران قرار می دهد و امکان ضبط و پخش آنها را فراهم می سازد.

سرویس ها

سرویس	وضعیت کنونی			
دانلود	• سرویس دانلود با قابلیت جستجوی انواع فایلها (نرم افزار، فیلم، صوت، کتاب و ...) در بیش از ۲۰ سایت معتبر دانلودی پیاده سازی شده است.			
بازار	• سرویس بازار با قابلیت جستجوی محصولات مختلف در بیش از ۲۰ سایت معتبر فروشگاه آنلاین و با نمایش مشخصات محصول (قیمت، موجودی، ویژگیها) پیاده سازی شده است.			
قیمت	• قیمت برخی از آیتم های شاخص (نظیر ارز، سکه، سهام بورس) در پارسی‌جو قابل جستجو و نمایش می‌باشد. • جستجو و نمایش قیمت طیف وسیعی از محصولات مهم (خودرو، لوازم خانگی و ...)			
موبایل	• تا اکنون، سه اپلیکیشن اندرویدی پارسی‌جو، خبرجو و ترجمه توسط پارسی‌جو ارائه شده است.			
نقشه	• اولین و تخصصی ترین سرویس نقشه بومی کشور در این مقیاس • ارائه بیش از ۴۵۰ لایه اطلاعاتی نقشه کل کشور • قابلیت ارائه سامانه مدیریت جامع مکانی			
علمی	• جستجو در یک میلیون مقاله انگلیسی و فارسی			
داغ	ورزشی	تقویم	اوقات شرعی	آب و هوا

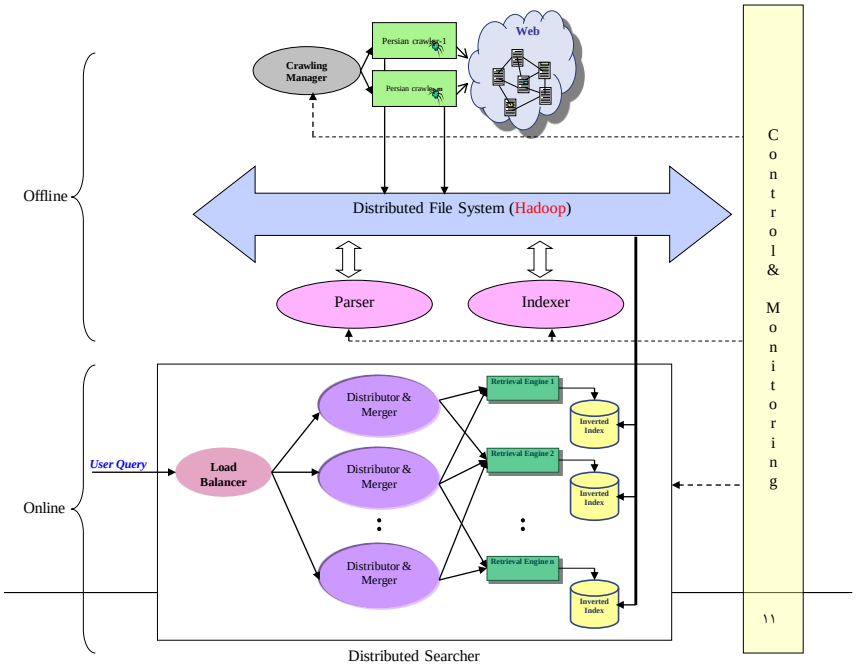
سرویس شاخص برای اندازه گیری کیفیت و ترافیک وب سایتها

نرخ بازدید پارسی جو

- آمار فعلی: روزانه ۱۶۰ هزار پرس و جو (بیش از ۱۰۰ هزار وب)
- تعداد کلیلک در وب: حدود ۱۰۰ هزار
- نرخ ماندگاری بالای کاربران
  - تعداد صفحات بازدید شده
  - زمان ماندن در سیستم بالای دوازده دقیقه

- تعدا پرس و جوی شهرها به ترتیب (بعد از تهران):
  - گرگان (۲۵۰۰)- ارومیه (۲۲۰۰)-یزد (۱۵۰۰)-تبریز (۱۴۰۰)-مشهد (۱۳۰۰)

معماری توزیع شده پارسی جو



آمار موجود داده های پارسی‌جو

- تعداد صفحات خزش شده یک و نیم میلیارد
- تعداد پیوندهای کشف شده بیش از ۸۰ میلیارد
- تعداد پیوندهای یکتا بیش از دو میلیارد
- تعداد واژهای یکتا بیش از میلیارد (یکی، دوتایی و سه تایی)

چالشهای موتور بومی

- تشخیص صفحات فارسی و صفحات تکراری
- بودجه بندی وب سایتها
- حدود ده میلیارد سند فارسی وجود دارد
- بخش خزش
- پردازش زبان
- صفحات اسپم
- بروز آوری صفحات

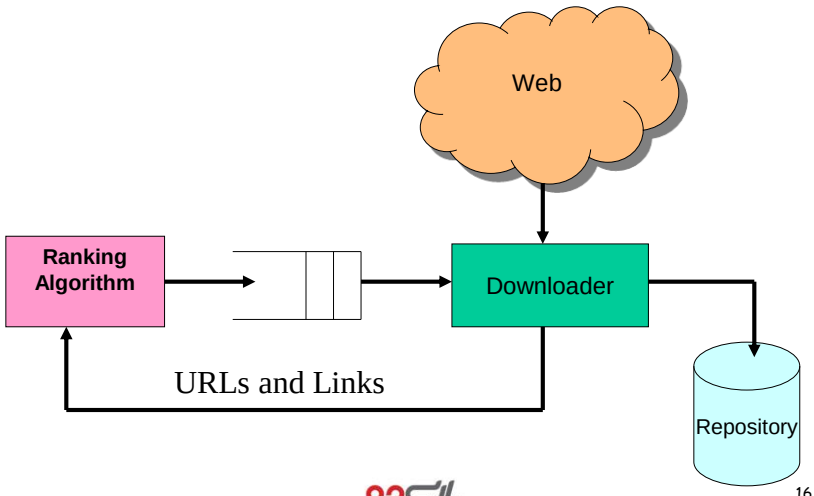
تشخیص صفحات فارسی

- هم پوشانی با زبان عربی و اردو
- بیش از ۵۰ درصد محتوا تکراری است
- بیش از ۲۰ درصد محتوای یک سایت تکراری است
- محتوای یکسان با آدرسهای مختلف

چالشهای خزش داده های حجیم

- بودجه بندی
- تشخیص پیوندهای یکتا
- رعایت ادب
- بازدید مجدد وب سایتها

معماری خزش



16

بودجه بندی

- هدف: خزش بهترین اسناد فارسی میباشد
  - میلیاردها صفحه فارسی
- مبتنی بر گراف
- مبتنی بر کلیک کاربر

تشخیص پیوندهای یکتا

- وجود بیش از ۸۰ میلیارد پیوند (که باید یکتایی آنها مشخص شود)
- نیاز به ساختمان داده کارا
- قدرت پاسخگویی هشتاد هزار آدرس در ثانیه
  - هر ثانیه دوهزارصفحه خزش میشوند و حدود هشتاد هزار پیوند استخراج میشود

رعایت ادب و بروز آوری

- برای رعایت ادب و تعادل در وب لازم است برای هر هاست صف جداگانه ای در نظر گرفته شود
  - بیش از بیست میلیون هاست وجود دارد
  - بیش از بیست میلیون صف!
- وب سایتها لازم است با توجه به زمان بروزآوری آنها خزش شوند
  - صفهای زمانی (ده دقیقه، نیم ساعت، یک ساعت و غیره)

چالش نمایه سازی

- با توجه به هدف با پاسخگویی به پرس و جوی طبیعی کاربران هیچ کدام از پایگاههای داده جوابگو نیستند
  - زبان طبیعی کاربر
  - عدم ساختار در صفحات وب
- استفاده از ساختار نمایه معکوس
- حجم زیاد نمایه ها
  - تقسیم به نمایه های کوچکتر(توزیع شده)
  - پارامترهای محتوایی از دقت کافی برخوردار نیستند

چالش سرعت جستجو

- **میانگین طول پرس و جو سه کلمه میباشد**
- **برای هر جستجوی سه کلمه ای بیش از ۱۵۰میلیون مقایسه و عملیات ریاضی لازم است**
  - **برای ۵۰ پرس و جو در ثانیه حدود ۷.۵ میلیارد عمل در ثانیه لازم است**
- **استفاده از حافظه پنهان هوشمند**
  - **با توجه به تکرار کلمات پرس و جو و نوع کلمات (مهم و غیر مهم)**

پردازش زبانی پرس‌وجوهای کاربران

- مراحل پردازش پرس‌وجو
  - پیش‌پردازش زبانی
    - همسان‌سازی نویسه‌ها
  - پیش پردازش‌های تغییردهنده کلمه
    - تصحیح املایی
    - ریشه‌یابی کلمات
  - تحلیل نحوی پرس‌وجو
    - تعیین نقش اجزا دستوری (مثلاً تعیین افعال)
  - تحلیل معنایی پرس‌وجو
    - تعیین حیطه موضوعی پرس‌و جو و دسته بندی آن

پیش پردازش‌های تغییردهنده کلمه

- پردازش‌های تغییردهنده کلمه، به آن دسته از پیش پردازش‌ها اطلاق می شود که با هدف بهبود شکل پرس‌وجو صورت می‌گیرند
  - تصحیح املایی و فاصله‌گذاری
    - مستلزم استفاده از اطلاعات آمار و مدل‌های زبانی مستخرج از وب
    - ترکیب مدل‌های زبانی و مدل کانال نویزی برای تصحیح مبتنی بر وب (تصحیح عبارت "حیاط موجودات زنده" به "حیات موجودات زنده")
  - ریشه‌یابی کلمات
    - کاهش تنوع کلمات جستجو با ریشه‌یابی کلمات و کاهش وندهای تصریفی مانند، تغییر "درختان میوه ایرانی" به "درخت میوه ایرانی" یا "درخت میوه ایران"

پردازش‌های نحوی پرس‌وجو

- تحلیل‌های نحوی پرس‌وجو، بیشتر از نوع پردازش‌های سطحی و نیمه عمیق هستند که به عنوان میان پردازش و با هدف بهبود پردازش‌های معنایی صورت می‌گیرند.
- تعیین افعال و کلمات توقف
  - در پرس‌وجوهای کاربران، تعداد کاربردهای مصادر بیشتر از افعال تصریف شده است.
  - تعیین کلمات توقف موجود در پرس‌وجوها
    - کلمات توقف در کاربرد وب محدود به کلمات توقف متداول نیستند و کلماتی مانند دانلود نیز همان ویژگی را دارند.
    - کلمات توقف در پرس‌وجوها حذف نمی شوند و اثر آنها تعدیل می‌شود.

پردازش‌های معنایی پرس‌وجو

- تحلیل‌های معنایی پرس‌وجو
  - تعیین کلمات کلیدی پرس‌وجو
    - کلماتی دارای بار معنایی بالا که منظور کاربر و نوع پرس‌وجو را بهتر مشخص مینمایند. (در عبارت "کدهای بازی GTA4" کلمه "کدهای" کلمه کلیدی این عبارت است که آنرا از سایر عبارت های جستجوی مربوط به بازی‌ها و بازی GTA4 به طور اخص، جدا می‌کند.)
  - تعیین دسته‌بندی پرس‌وجو
    - پرس‌وجوهای کاربران می‌توانند در سه دسته اطلاعاتی، پیمایش وب، عملکردی دسته‌بندی شوند.
    - هر یک از این دسته‌ها دارای کلمات کلیدی و اسلوب متفاوتی بوده و برخورد با آنها در نمایش نتایج نیز متفاوت است.

پردازش گراف وب

- **گراف با یک و نیم میلیارد گره و بیش از ۸۰ میلیارد یال**
- **اجرای الگوریتم های رتبه بندی مبتنی بر گراف (محبوبیت)**
- **اجرای الگوریتم های تشخیص صفحات اسپم**

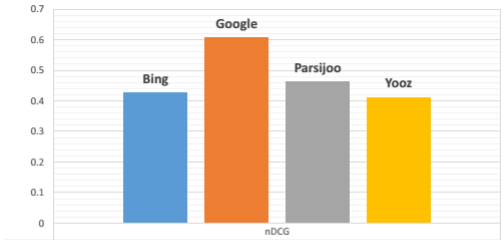
تشخیص صفحات اسپم

- بیش از ۵۰ درصد صفحات اسپم هستند
- روشهای مبتنی بر اتصال
- روشهای مبتنی بر محتوا
- روشهای ترکیبی

بروز آوری اسناد

- روزانه بیش از ده میلیون سند خزش میشود
  - ۵ میلیون جدید
  - ۵ میلیون بازدید مجدد
- هدف رساندن زمان بروزآوری به زیر یک ساعت
  - نمایه سازی
  - گراف وب
- در حال حاضر زمان بروز آوری ۴۸ ساعت میباشد

ارزیابی موتورهای جستجوی داخلی و خارجی توسط مرکز تحقیقات مخابرات (۱۴۰کاربر و ۴۰۰پرس و جو)



	nDCG
Bing	0.427272166
Google	0.608898327
Parsijoo	0.46336862
Yooz	0.413139052

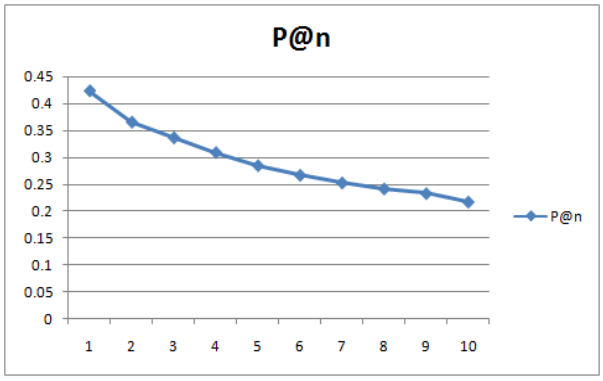
چالش افزایش دقت

- رتبه بندی هوشمند و کارا
  - ویژگیهایی محتوایی و اتصالی
- استفاده از رفتار کاربر
  - پرس و جوها و کلیکهای روی پیوندها
  - مدل کلیک از گذرداده

مقایسه با گوگل

تاریخ تست	P@1	P@2	تعداد پرس و جو
۱۲/۱۱/۹۱	۲۵/۰	۱۸/۰	۱۳۰۰
۱۷/۱۲/۹۱	۴۱/۰	۲۹/۰	۳۸۰۰
۲۱/۵/۹۲	۴۹/۰	۴۳/۰	۲۰۰۰
۸/۸/۹۲	۶۲/۰	۵۹/۰	۱۳۵۰۰
۶/۱۰/۹۳	۶۳/۰	۶۰/۰	۱۳۵۰۰
۲۰/۲/۹۴	۶۴/۰	۶۲/۰	۱۳۵۰۰
۲۰/۸/۹۴	۶۸/۰	۶۵/۰	۱۳۵۰۰
۱۷/۱/۹۵	۶۸/۰	۶۴/۰	۱۳۵۰۰
۲۰/۴/۹۵	۶۹/۰	۶۴/۰	۱۳۵۰۰

مقایسه با گوگل (پرس و جوهای اطلاعاتی)



چالش جذب کاربر

- کیفیت نتایج
- سرویس های بومی و محلی
- دسترسی به داده های محلی

سپاس از حسن توجه شما