اهداف و روند اجرای پروژه طراحی سامانه رتبه بندی

مربوط به پروژه‌ی تحلیل نیازمندی‌ها، طراحی و پیاده‌سازی سامانه بومی رتبه بندی و پایش وب‌سایت‌ها

ارائه‌دهنده: لیلا ربیعی
پژوهشکده: **فناوری ارتباطات**
گروه: **مدیریت یکپارچه شبکه**
تاریخ: **هجدهم خرداد ۱۳۹۵**

www.itrc.ac.ir

مفاهیم کلی رتبه بند

رتبه بند چیست؟ رتبه‌بندی وبگاه‌ها راهکاری است که تعیین می‌کند چه وبگاه‌هایی از دید کاربران محبوب‌تر هستند.

یک مکانیسم برای ارزش‌دهی به هر سایت با توجه به معیارهای اندازه‌گیری شده

دلایل نیاز به رتبه بند بومی؟ الکسا مرجعی نه چندان دقیق برای رتبه‌بندی وب‌گاه‌ها در ایران و حتی در سطح بین المللی

- عدم دسترسی الکسا به منابع داده ای ایرانی
- وجود راه‌کارهایی برای فریب در رتبه الکسا
- تأثیر ترافیک غیر واقعی بر رتبه الکسا
- غیرمتوازن بودن داده‌های جمع‌آوری شده از وب‌سایت‌هاست

نیاز به یک رتبه‌بند که علاوه بر حل مشکلات مطرح شده بتواند با تکیه بر منابع داده‌ای داخلی آمار دقیق و قابل استنادی ارائه کند و بتواند اعتماد کاربر داخلی را چه در سطح دولت و چه در سطح صاحبان کسب و کار جلب کند

هدف از ایجاد رتبه بند بومی

ایجاد جایگزین بومی مناسبی برای الکسا و امثال آن در کشور که ضمن تکیه بر منابع داده ای قابل اعتمادتر، نقاط ضعف الکسا را مرتفع نماید

اهداف کلان:

منظور از تحلیل وب، اندازه گیری،جمع آوری، پردازش و گزارش اطلاعات وب با هدف درک و بهینه سازی کاربردهای آن برای صاحبان وب سایت ها می باشد. برخی از معیارهای تحلیل وب می توانند به منظور افزایش دقت رتبه تخصیص داده شده به وب سایت ها مورد استفاده قرار گیرند. به عنوان مثال تحلیل هایی که منجر به شناسایی ترافیک های غیر واقعی می گردد.

سامانه بومی رتبه بندی و تحلیل وب سایت ها

رتبه بندهای بومی مطرح در دنیا

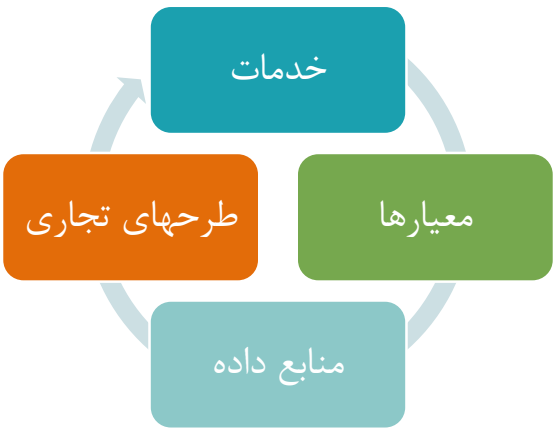
نرم‌افزار یا سامانه بومی: نرم‌افزاری که در یک کشور توسط متخصصین آن کشور با اولویت دهی به نیازها و سیاست‌های بومی برای آن نرم‌افزار، با بهره گیری از زیرساختهای موجود در کشور، ایجاد می‌شود

مبتنی بر ترافیک	نرم‌افزارهایی است که علاوه بر ارائه تحلیل از وب‌سایت‌ها آنها را در زمینه‌های مختلف(دسته بندی جغرافیایی، دسته‌بندی محتوا و ...) رتبه‌بندی می‌کنند
مبتنی بر ارجاعات	نرم‌افزارها و یا سرویس‌های تحت وبی هستند که وب‌سایت‌ها را بر مبنای گراف روابط صفحاتشان ، مشابه رتبه‌بندی جویشگران، رتبه‌بندی می‌کنند
سیستم‌های ترکیبی	وب‌سایت‌ها و سرویس‌های تحت وبی است که رتبه وب‌سایت متقاضی را در رتبه-بندهای مطرح جهانی در هر دو دسته مبتنی بر گراف و مبتنی بر آمار ترافیکی نشان داده و همچنین رتبه احتمالی یک وب‌سایت را در جویشگران مختلف و مطرح، جهت تحلیل‌های SEO، نشان می‌دهد
مبتنی بر رأی گیری	سرویس‌های تحت وب و نرم‌افزارهایی است که وب‌سایت‌ها را توسط معرفی خودشان و یا افراد شناسایی می‌کنند و مبتنی بر نظرات مردم و سایر معیارها به آن وب‌سایت‌ها رتبه می‌دهند
رتبه‌بندهای جعبه‌سیاه	وب‌سایت‌ها و سرویس‌های تحت وبی هستند که یک لیست از برترین سایت‌ها را بدون هیچ توضیحی از معیارها و چگونگی آن در انظار عموم می‌گذارند

رتبه بندهای بومی مطرح در دنیا

ردی ف	نام رتبه‌بند	کشور ایجاد کننده	نوع رتبه‌بند
۱	Alexa	آمریکا	مبتنی بر ترافیک
۲	Comscore	آمریکا	مبتنی بر ترافیک
۳	Quantcast	آمریکا	مبتنی بر ترافیک
۴	Similar web	انگلستان	مبتنی بر ترافیک
۵	Google Page Rank	آمریکا	مبتنی بر ارجاعات
۶	Advanced Web Ranking	آمریکا	ترکیبی
۷	Generate it	آمریکا	ترکیبی
۸	Top-Site-List	آمریکا	مبتنی بر رأی گیری
۹	Site-Rank-Data	آمریکا	ترکیبی و مبتنی بر رأی گیری
۱۰	MOZ	آمریکا	جعبه سیاه
۱۱	VALBOT	آمریکا	جعبه سیاه
۱۲	Chinesetop100	چین	جعبه سیاه
۱۳	Compete	آمریکا	مبتنی بر ترافیک

نرم افزارهای رایج تحلیل و رتبه بندی



- Alexa
- comScore
- Google Analytics
- Google page rank

جایگاه الکسا

- الکسا مرجعی **نه چندان دقیق** برای رتبه‌بندی سایتها در سطح محلی و بین المللی است. مدیران سایتها همواره به دنبال افزایش این رتبه هستند و راه های زیادی برای **دور زدن رتبه الکسا** به وجود آمده است.

- با انجام این پروژه و راه اندازی سرویس بومی رتبه بندی وبسایتها، نقاط ضعف الکسا شناسایی شده و **راه کارهایی بر اساس داده های دقیقتر و مستندتر** برای رتبه بندی، پیاده سازی خواهند شد. در این پروژه سعی بر آن است که با توجه به اصالت در اطلاعات، حفظ امنیت و سرمایه مصرف کنندگان و قابلیت اعتماد که همگی با پشتوانه بومی بودن محصول بوجود خواهند آمد، جایگزین مناسبی برای الکسا و امثال آن در کشور ایجاد شود

الکسا در کنار سایر رتبه بندها

- در این پروژه الکسا در کنار سایر نرم‌افزارهای رایج تحلیل و رتبه‌بندی از جمله **گوگل آنالیتیکس** و **کام اسکور** مورد بررسی قرار گرفته است.

- الگوریتم رتبه‌بندی سامانه بومی به هیچ وجه منطبق بر الگوریتم رتبه‌بندی الکسا نیست.

- هرچند هدف این پروژه ارائه راهکاری با حداکثر جلوگیری از امکان تقلب است اما پیاده سازی **راهکارهای جلوگیری از تقلب** اجتناب ناپذیر است.

سرویس‌های ایرانی آنالیز لاگ

- سرویس شرکت بیان
  - **Alexa Pro** : این محصول در واقع با استفاده از یک اسکریپت به جمع آوری داده ها پرداخته است و سپس بر اساس همین داده‌ها، رتبه بندی و تحلیل انجام میدهد که این روش در واقع بخش خیلی جزئی ازکار سامانه رتبه بندی میباشد.
- پرشین استت
  - پرشین استت با موافقت صاحب یک وبسایت به راحتی میتواند با قراردادن یک قطعه کد کلیه فعالیت های صورت گرفته در وب سایت را رصد کند و آمارها و تحلیلهای مورد نیاز را به کاربر ارائه کند. اگرچه با موافقت کاربران همواره یک لیست از وبسایتهای برتر نمایش داده میشود اما مشکل اساسی این است که تنها وب سایت های ثبت شده در این رتبه بندی شرکت داده می شوند. در نتیجه به علت کوچک و نا متوازن بودن جامعه آماری نمی توان پرشین استت را بک عنوان یک سیستم رتبه بندی لحاظ کرد.

سرویس‌های ایرانی آنالیز لاگ

- مرور
  - سرویس مرور بسیار شبیه سرویسی است که پرشین استت ارائه کرده است. یعنی برای فعال سازی محصول و خدمتی که ارائه می شود لزوما نیاز به قرارگیری کد اسکریپت در وب سایت مورد نظر کاربر است. در نتیجه محدوده سرویس دهی آن تنها محدود به تعدادی وب سایت است که به صورت داوطلبانه خود را در مجموعه وب سایت های این تحلیلگر ثبت نموده اند.
- پروژه مرکز رسانه های دیجیتال
  - این مرکز به دلیل **جایگاه حاکمیتی** و اعتماد سازی که در میان صاحبان وب سایت ها ایجاد نموده است میتواند به صورت بالقوه، یکی از منابع خوب جمع آوری اطلاعات از وب سایت های داخلی و همچنین مرجع مناسبی **برای اطلاع رسانی و تبلیغ** محصول رتبه بند بومی باشد.

بکارگیری نرم افزارهای Open source

- ابزار open sourceی که پاسخگوی **همه اهداف پروژه** باشد، وجود ندارد. البته ابزارها و کتابخانه‌هایی هستند که می‌توانند برای قسمت‌های مختلف پروژه استفاده گردند که در این پروژه نیز از آن‌ها استفاده خواهیم کرد، ولی هیچ یک از این ابزارها مستقلا برای کارایی مدنظر پروژه مناسب نیستند. به طور مشابه، در مورد سایر پروژه‌های بومی هم ابزارهای اوپن سورس وجود دارند

  - ابزارهای اوپن‌سورس راه‌اندازی وبلاگ مانند
    - Apache Roller
  - موتورهای جستجوی اوپن سورس مانند
    - Solr + Nutch
  - مترجم‌های ماشینی مانند
    - Marie و Joshua و Giza++ و Apertium و Matxin و OpenLogos و Moses و Anusaaraka.

بکارگیری نرم افزارهای مدیریت لاگ

- برخی نرم افزارهای مدیریت لاگ منبع باز:
  - AWstat
  - Graylog
  - Logcheck
  - Logwatch
  - Logstash
  - ...

❋ با هدف مانیتور کردن و رفع عیب سریع تر ایجاد شده اند

❋در حوزه **مدیریت** و administration سیستم فعال هستند

❋ در رده **سازمانی** فعالیت میکنند و برای سامانه بومی با محدوده وسیع ناکارآمد هستند

بکارگیری Google Analytics

- ارائه اطلاعات بازدید توسط خود وبسایتها (به صورت داوطلبانه)، این امکان با دو سناریو در پروژه پیش‌بینی شده است:

۱) تعامل با یکی از نهادهای مسئول که چنین سامانه و اطلاعاتی را در دست دارد

۲) ارائه سرویس APIآنالیتکس به صورت مستقل از سایر سیستم‌ها

- ارائه آمار مستخرج از API گوگل انلیتیکس به یک مرجع دارنده رتبه و لیست هزار وب سایت کشور

چنین مرجعی وجود ندارد، این پروژه در آینده نقش این مرجع را ایفا خواهد کرد و هدف از تعریف پروژه ایجاد چنین مرجعی است

منابع داده‌ای

- از مهمترین بخشهای پروژه جمع آوری داده‌های تحلیل ترافیکی وبسایتها است و اساسا ابزار تحلیلگر وب را نمی‌توان بدون در دست داشتن داده ساخت.
- سیستم بدون وجود داده طراحی نشده، و توسعه سیستم تا زمان تهیه‌ی همه‌ی انواع داده به تعویق نخواهد افتاد و دستیابی به برخی منابع داده نیاز به امکان سنجی دارد.
- انواع منابع داده‌ای لیست شده:
  - منابع داده ای برون وبگاه (Off Site) که نهادهایی صاحب آن داده‌ها هستند و بر اثر تعاملات و توافقات دسترس پذیر خواهند شد.
  - منابع داده‌ای بر وبگاه (On Site) مانند کوکی ، نوارابزار، اسکریپت و API

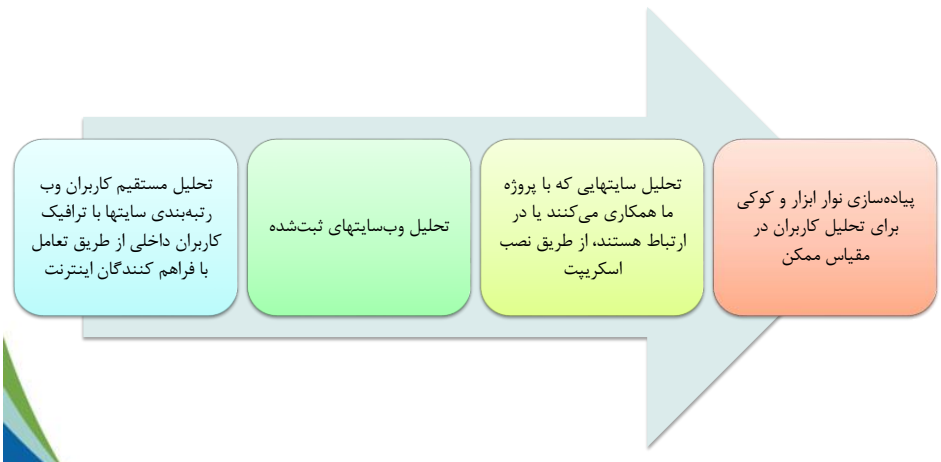
شیوه فنی پیاده‌سازی

❖ **از آنجایی که مهم‌ترین مزیت رقابتی این محصول، امکان استفاده و «تجمیع» منابع داده‌ای مختلف برای رسیدن به دقت و اعتبار لازم است، شیوه فنی طراحی و پیاده‌سازی منابع داده‌ای به عنوان *یکی از فعالیتهای پروژه در گامهای طراحی و پیاده سازی* است که مطابق پیشنهاد پروژه شامل موارد زیر است:**

- بررسی مسائل فنی و حقوقی مرتبط با دریافت داده از منابع مربوط و ارائه راهکارها و گزینه های مختلف ممکن برای همکاری با دارندگان داده
- بررسی و ارائه راهکارهای فنی برای استفاده همزمان از چند منبع داده ای برای بهبود عملکرد رتبه بندی، و ارائه راهکار برای ادامه روند عملیات سیستم در صورتی که یکی از منابع داده ای بطور موقت در دسترس نباشد.
- بررسی و ارائه راهکارهای فنی برای مدیریت حجم بالای داده ها
- ارائه راهکار برای نگهداری و پردازش داده های جریانی (مانند داده های ترافیکی دریافتی از طریق اسکریپتها یا افزونه ها به مرورگرهای کاربران)
- ارائه راهکار برای کاهش احتمال تقلب برای افزایش رتبه توسط دارندگان تارنماها
- ارائه راهکار برای محاسبه بهتر و عادلانه تر رتبه هر سایت
- ارائه تحلیلهای ترافیکی مختلف متناسب با نیاز هر یک از ذینفعان در صورت نیاز

15

منابع داده‌ای مورد استفاده با اولویت میزان پوشش‌دهی داده های مورد نیاز



16

بررسی جنبه‌های حقوقی پروژه

- **جنبه‌های حقوقی بررسی شده :**

۱. مفهوم رتبه‌بندی
۲. هدف رتبه‌بندی
۳. موضوع رتبه‌بندی
٤. مبنای رتبه‌بندی
٥. گستره رتبه بندی
٦. ابزار رتبه بندی
۷. منابع رتبه‌بندی
۸. کارگزار(ان) رتبه‌بندی
۹. بهره‌برداران رتبه‌بندی
۱۰. سوژه‌های رتبه‌بندی
۱۱. راهبر(ان) رتبه بندی
۱۲. قواعد رتبه بندی
۱۲-۱. مالکیت
۱۲-۱-۱. مالیت داشتن داده ها
۱۲-۱-۲. اهلیت دارنده داده ها
۱۲-۲. صلاحیت
۱۲-۳. تمامیت
۱۲-٤. حیثیت
۱۲-٥. امنیت
۱۳. حق ها/امتیازهای برآمده از رتبه بندی
۱٤.جمع‌بندی و نتیجه‌گیری
الف) تمهیدهای حقوقی کوتاه مدت
ب) تمهیدهای حقوقی میان مدت
پ) تمهیدهای حقوقی دراز مدت

17

آورده‌های جنبی پروژه علاوه بر سامانه بومی تحلیل و رتبه‌بندی وب‌سایت‌ها

- تجربه اجرای یک پروژه با چالشهای مختلف فنی و حقوقی
- فراهم شدن بستری برای نگهداری و مدیریت و پردازش حجم بالای داده های ترافیکی برای انجام پروژه های بسیار متنوع بعدی که نیاز به داده ترافیکی دارند
- فتح باب و رفع موانع فنی و حقوقی (یا حداقل با مشکلات و طرح راهکارهای ممکن) برای بکارگیری منابع عظیم داده ای موجود در زیرساخت و ISPها و ... برای پروژه های بعدی. واضح است که هر شرکت دانش بنیانی هم برای استفاده از این منابع داده ای با مشکل روبرو است و یکی از آورده های این پروژه میتواند ارائه راهکار برای این مشکل و یا حداقل توجه دادن مراجع قانونی به نیازها و مشکلات در این حوزه جهت رفع موانع در آینده است.

18

حوزه کار سامانه در ارزیابی وبسایت‌ها

19

دلایل ارائه یک عدد به عنوان رتبه

- مراجعه مخاطب : کاربران وب از میان اقشار مختلف جامعه هستند و جهت درک میزان اعتبار و شهرت یک وبسایت به آن استناد می کنند.
- ارائه تحلیلهای مختلف برای مخاطب نامبرده موارد زیر را همراه دارد:
  - ✓ پیچیدگی
  - ✓ زمانبر بودن
  - ✓ عدم گنجایش در حوصله ایشان
  - ✓ نبود دانش کافی و احتمال برداشت های غلط از تحلیلهایی مانند نرخ خروج و بازدید صفحه یکتا و ...
- لحاظ نمودن کارشناسانه معیارها با وزن دهی مناسب برای نتیجه گیری از آنها در قالب عدد رتبه
- نیاز به عدد رتبه برای ایجاد مرجع معتبری برای لیست نمودن وبسایتهای یک تا ۱۰۰۰ در کشور
- ارائه رتبه به وبسایتها پیشینه دار بوده و قبلا توسط شرکتهای بزرگی چون Google و ComScore ارائه شده است.

20

دلایل انجام پروژه در ارگان دولتی و نه شرکتهای خصوصی

- هدف این پروژه، جاسوسی اطلاعات نیست بلکه تحلیل داده‌ها با حفظ حریم خصوصی و با رعایت جنبه‌های حقوقی بحث است.
- شرکت‌های بزرگ دنیای فناوری اطلاعات با همین سازوکار با دولتها همکاری کرده و اطلاعات مورد نیاز دولت‌ها را در اختیار آن‌ها قرار می‌دهند، بدون این که حریم خصوصی کاربران نقض شود.
- بحث اعتمادسازی و حل چالش‌های حقوقی در این زمینه برای نهاد حاکمیتی (نسبت به بخش خصوصی) ساده‌تر صورت می‌گیرد.
- فراهم شدن بستری برای نگهداری و مدیریت و پردازش حجم بالای داده های ترافیکی که هزینه بالایی را می طلبد، برای نهادهای دولتی با تضمین بالاتری صورت می‌پذیرد.

21

دلایل انجام پروژه در ارگان دولتی و نه شرکتهای خصوصی

- امنیت ذخیره و نگه داری داده‌های ترافیکی کشور که از ارزش بالایی برخوردار است در نهادهای دولتی با ضریب اطمینان بالاتر و نظارت بهتری حفظ می‌شود.
- از مهمترین دستاوردهای ارائه رتبه به وبسایتها تاثیر آن در کسب و کار ایشان است که در اینصورت احتمال تقلب در ارائه رتبه در نهادهای دولتی به دلیل بی طرف بودن ایشان بسیار پایین تر از شرکتهای خصوصی است.
- مداومت و پایداری سامانه وابسته به مداومت مالکیت آن است و با عنایت به اینکه تضمینی برای پایداری شرکتها وجود ندارد، مالکیت سامانه با نهادهای حاکمیتی ارجحیت دارد.
- هزینه تخمینی با محاسبه بر اساس اصول تعریف پروژه در پژوهشگاه ۲ ملیارد تومان بوده است.

22

وزارت ارتباطات و فناوری اطلاعات
پژوهشگاه ارتباطات و فناوری اطلاعات
(مرکز تحقیقات مخابرات ایران)
ITRC

با تشکر از توجه شما