پیاده سازی موتورجستجو برای زبان فارسی

توسط : رضا شیرازی مفرد

www.rezashirazi.com
info@rezashirazi.com

فهرست مطالب

- مقدمه
  - اهمیت موتورهای جستجو
  - اجزای موتور جستجو
  - انواع الگوریتم های رتبه بندی
  - موتورهای جستجوی متن باز
- موتور جستجوی فارسی
  - مشکلات موجود در زبان فارسی
  - پیاده سازی خزنده وب
  - شناسایی صفحات فریب آمیز
  - طراحی یک سیستم خبره فازی برای رتبه بندی درصد مناسب بودن صفحات
  - الگوریتم های رتبه بندی در موتورجستجوی فارسی وب

اهمیت موتورهای جستجو – آمار و اطلاعات

- بیش از ۸۰% ترافیک اینترنت از طریق موتورهای جستجوی تامین می شود.
- با توجه به رشد زیاد اطلاعات اینترنت و نیاز کاربران در زمینه دستیابی به اطلاعات، الگوریتم های رتبه بندی صفحات اینترنتی در موتورهای جستجو از اهمیت بالایی برخوردار هستند.
- رتبه بندی صفحات در موتور جستجو تا حدود زیادی وابسته به ویژگی های زبانشناسی و فرهنگی است بنابراین نتایج یک موتور جستجوی بومی سازی شده در صورتی که از الگوریتم های مناسب رتبه بندی استفاده کند و اطلاعات مناسبی داشته باشد، می تواند بهتر از گوگل عمل کند.
- برخی از کشورهایی که به سمت بومی سازی موتورهای جستجو رفته اند و موفق بوده اند: روسیه، چین، هند، جمهوری چک و کره جنوبی.

آمار استفاده از موتورهای جستجو در کشورهای مختلف

	Search Engines			
Country	Leader	Share	Runner Up	Share
Argentina	Google	95%	Bing	4%
Australia	Google	87%	Bing	3%
Brazil	Google	97%	Bing	2%
Canada	Google	78%	Bing	6%
Czech Republic	Seznam	45%	Google	45%
China	Baidu	76%	Google	22%
Denmark	Google	97%	Bing	2%
Egypt	Google	95%	Yahoo/Bing	5%
Finland	Google	95%	Other	5%
France	Google	92%	Bing	4%
Germany	Google	89%	t-online	3%
Hong Kong	Yahoo	N/A	Google	N/A
Japan	Yahoo Japan*	56%	Google	31%
Malaysia	Google	92%	Yahoo	5%
Mexico	Google	91%	Bing	7%
The Netherlands	Google	94%	Vinden	3%
New Zealand	Google	93%	Bing	2%
Philippines	Google	N/A	Yahoo	N/A
Poland	Google	97%	Other	3%
Russia	Yandex	62%	Google	26%
Saudi Arabia	Google	97%	Yahoo/Bing	3%
Slovakia	Google	99%	Yahoo/Bing	1%
South Korea	Naver	73%	Daum	18%
Spain	Google	96%	Yahoo/Bing	4%
Turkey	Google	N/A	Yandex	N/A
United Kingdom / UK	Google	94%	Bing	5%
United States	Google	72%	Yahoo	14%

منبع:
http://returnonnow.com/2012/06/search-engine-market-share-country/

ویژگیهای مورد انتظار از یک موتور جستجو

- سرعت پاسخ دهی مناسب
- نتایج قابل قبول و مناسب
- رتبه بندی صفحات بر اساس انتظار کاربران:
  - از لحاظ مکانی (مثال: نتایج مورد انتظار کاربر پس از جستجوی عبارت پیتزا فروشی)
  - از لحاظ مفهومی (مثال: ببر، شاهین)
  - عبارات مترادف (مثال: ویزای دانشجویی، اقامت دانشجویی، پذیرش تحصیلی، مهاجرت دانشجویی و ...)
- عدم نمایش صفحات فریب آمیز

ویژگیهای مورد انتظار از یک موتور جستجو – ادامه

- رتبه بندی بر اساس موضوع وب سایت و نه صرفا اطلاعات صفحات (چگالی اطلاعات درون سایت)
- عدم نمایش صفحات مختلف از یک وب سایت در نتایج
- عدم نمایش صفحات خراب
- دارا بودن بخش ابزار مدیران سایت
- تغییر وزن پارامترها در دوره های زمانی کوتاه برای جلوگیری از دستکاری نتایج توسط سایت های فریبکار

ویژگیهای مورد انتظار از یک موتور جستجو– ادامه

از آنجائی که مدیران سایت های اینترنتی برای افزایش درآمد خود به دنبال هدایت بازدیدکننده به سایت های مورد نظرشان هستند و برای این کار در جهت بالا بردن رتبه خود در موتورهای جستجو حرکت می کنند، راه های مقابله با اینگونه صفحات که به اصطلاح آنها را صفحات فریب آمیز می نامیم نیز از اهمیت بسیار بالائی برخوردار است.

لذا همواره جنگی میان موتورجستجو و مدیران سایت های اسپم در جریان است و باید پارامترهایی را در جهت جلوگیری از دستیابی به رتبه های بالا توسط اینگونه سایت ها در نظر بگیریم تا نتایج قابل قبولی داشته باشیم.

برخی از تکنیکهای سیاه

برخی از تکنیکهای سیاه سئو که موتور جستجو باید از آنها جلوگیری کند عبارتند از:

- Cloacking
- GEO Targeting
- Reapeted keyword
- Duplicate Data
- Hidden Data
- IP Delivery
- Googling
- Link Building:
  - Hidden Link Building
  - Buying Link
  - Link schemas

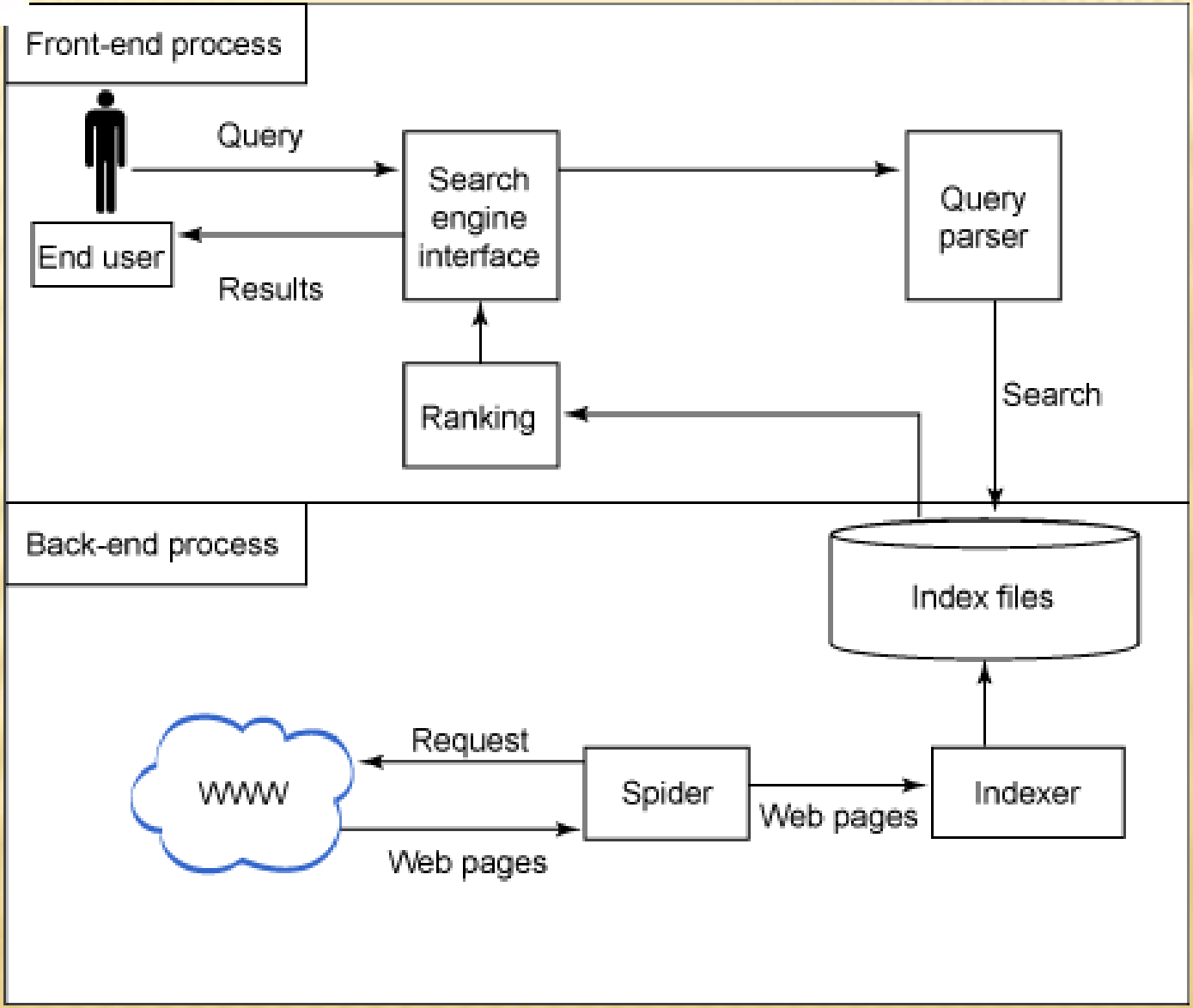
نحوه اختصاصی سازی نتایج توسط گوگل

Google Translate BETA

Google Patent Search BETA

مت کاتز: بهترین نتیجه جستجو برای شما، نتیجه ای است که دوستان شما آن را پیشنهاد داده باشند.

اجزای موتورجستجو



موتورهای جستجوی کدباز

- Apache Solr
- BaseX
- Clusterpoint Server (freeware licence for a single-server)
- DataparkSearch
- Ferret
- Hyper Estraier
- KinoSearch
- Lemur/Indri
- Lucene
- mnoGoSearch
- Sphinx
- Swish-e

موتورهای جستجوی کدباز – مقایسه

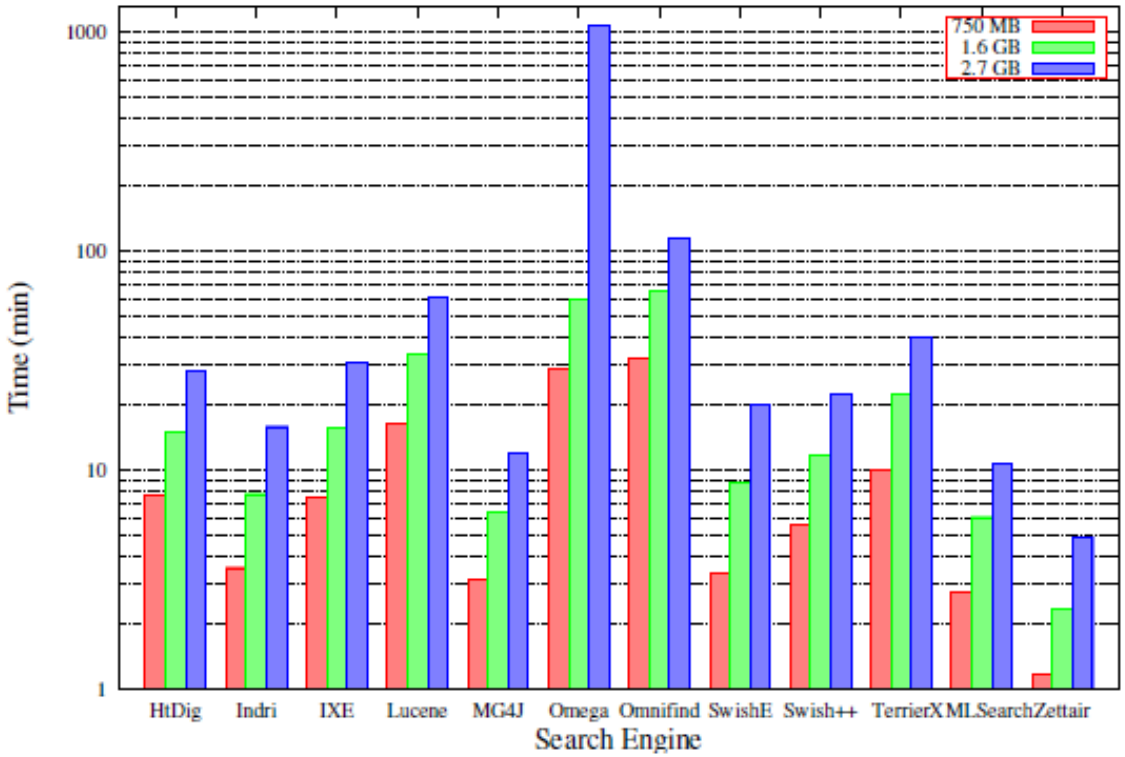


Figure 5.1: Indexing time for document collections of different sizes (750MB, 1.6GB, and 2.7GB) of the search engines that were capable of indexing all the document collections.

http://wrg.upf.edu/WRG/dctos/Middleton-Baeza.pdf

رتبه بندی صفحات وب

- Yahoo: http://learningtorankchallenge.yahoo.com/
- Microsoft LETOR:
  - http://research.microsoft.com/en-us/um/beijing/projects/letor//
- Learning to Rank:
  - http://en.wikipedia.org/wiki/Learning_to_rank

الگوریتم های رتبه بندی

- روش های مبتنی بر پیوند
  - الگوریتم Page Rank
  - الگوریتم HITS
  - الگوریتم SALSA
- روش های مبتنی بر محتوا
  - الگوریتم TF-IDF
  - الگوریتم BM25
- روش های ترکیبی
  - الگوریتم ComRank

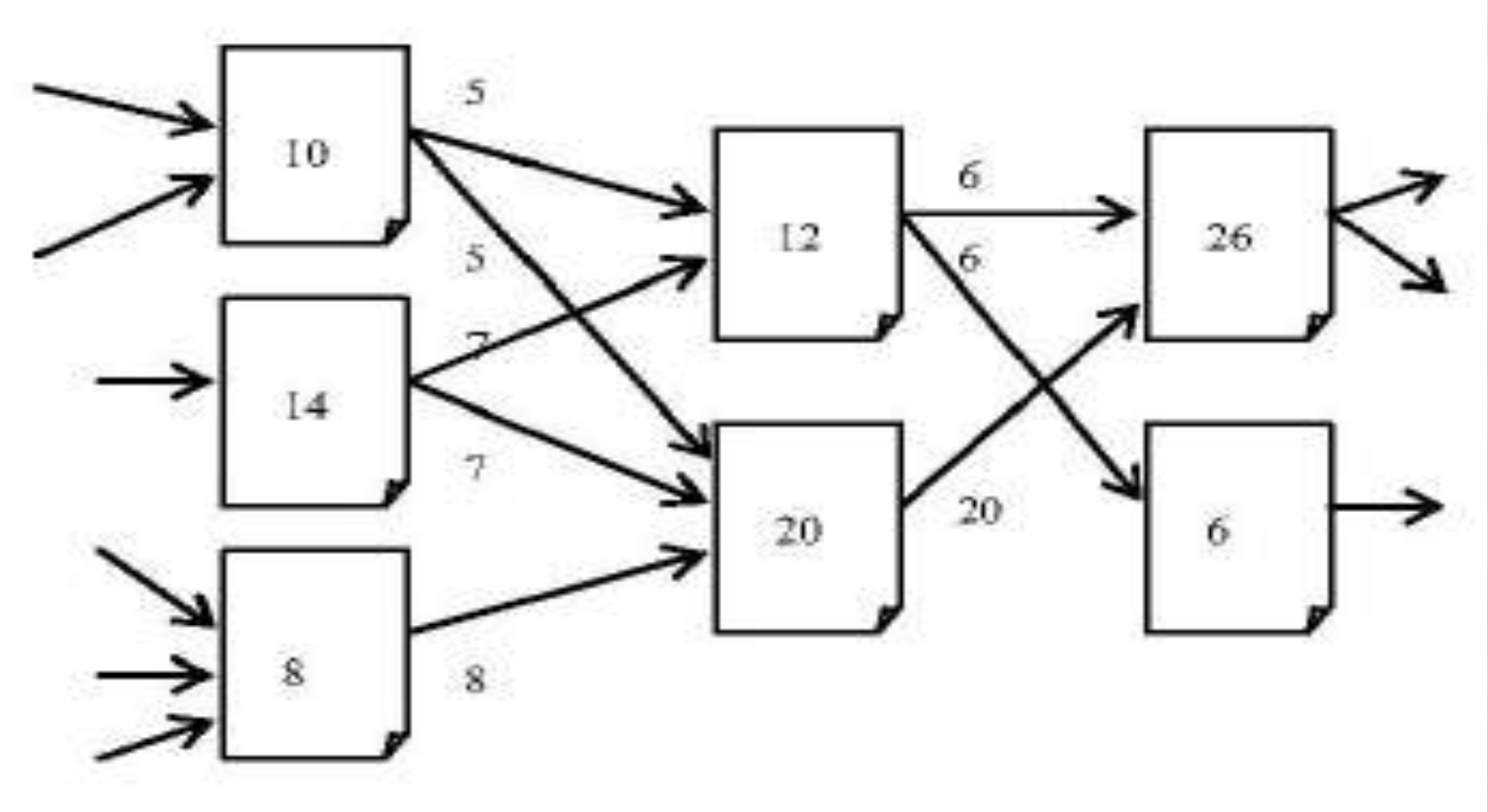
الگوریتم رتبه صفحه

- رتبه صفحه یا Page Rank عددی است بین ۰ تا ۱۰ که از طرف گوگل به صفحات اینترنتی نسبت داده می شود. هر چه میزان این عدد بیشتر باشد نشان دهنده این موضوع است که صفحه فوق از اهمیت بیشتری برخوردار است.
- ایده فوق توسط بنیانگذاران گوگل مطرح شد و برگرفته از معیار citation در مقالات علمی است.
- رابطه محاسبه PR به صورت زیر است :

$$R(u) = c \sum_{u \in B_u} \frac{R(u)}{N_u}$$

رتبه هر صفحه بر اساس مجموع رنک لینک های ورودی تقسیم بر تعداد لینک خروجی آنها با ضریب نرمال سازی C محاسبه می شود.

الگوریتم رتبه صفحه – مثال



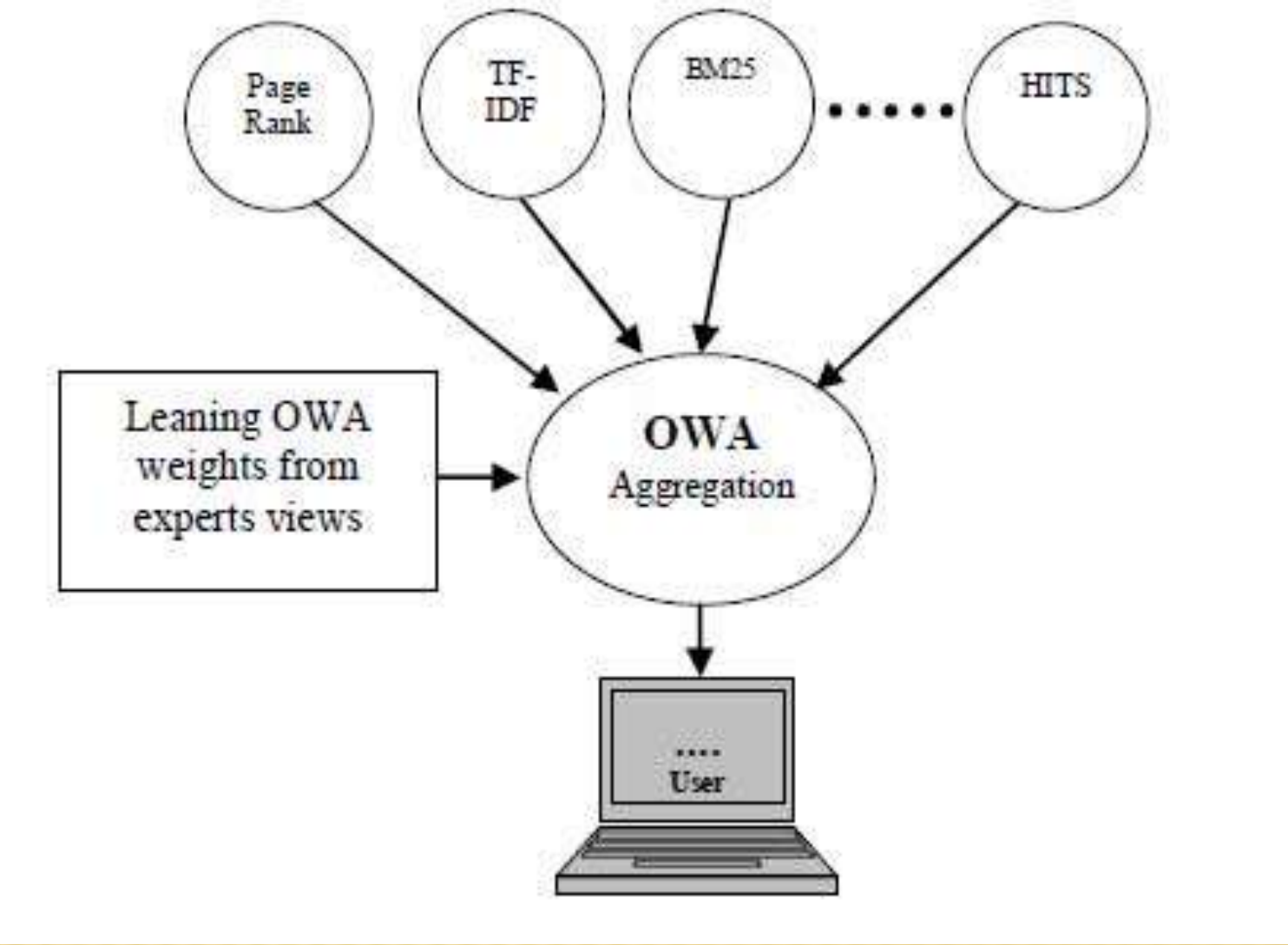
الگوریتم رتبه صفحه – ادامه

- مهمترین مزیت الگوریتم Page Rank پیدا کردن نتیجه بر اساس کلمات مترادف است. به دلیل اینکه سایت های مختلف با عبارات مختلف به یک سایت پیوند می دهند.
- مشکل اصلی الگوریتم Page Rank، غنی تر شدن اغنیا است. به این معنا که اگر یک سایت صفحاتی که رتبه بالایی دارند داشته باشد سایر صفحات آن سایت نیز رتبه بالایی خواهند داشت و با گذشت هر مرحله از محاسبه الگوریتم رتبه صفحات این سایت افزایش خواهد یافت.
- الگوریتم های HITS و SALSA توسعه یافته الگوریتم Page Rank هستند.
- با وجود اینکه دو الگوریتم HITS و SALSA سعی در برطرف کردن مشکلات Page Rank دارند، اما نتایج بدست آمده نشان می دهد که بهبود چندانی در نتایج جستجو توسط آنها نسبت به الگوریتم Page Rankحاصل نمی شود.

الگوریتم ترکیبی COMRANK

- الگوریتم ComRank، بر اساس ترکیب الگوریتم های مبتنی بر پیوند و الگوریتم های مبتنی بر محتوا که بر اساس چگالی اطلاعات صفحه عمل می کنند، رتبه بندی نتایج را انجام می دهد.
- این الگوریتم بر اساس عملگر OWA (میانگین وزن ترتیبی) کار می کند. مراحل این الگوریتم به صورت زیر است:
- فاز یادگیری:
  - افراد خبره
  - الگوریتم های یادگیری
- فاز رتبه بندی: پس از ارسال پرسش Q ، رتبه بندی در الگوریتم های مختلف انجام شده و نتایج بر اساس وزن ها با هم ترکیب می شوند و نتیجه رتبه بندی بر اساس نتایج بدست آمده به کاربر نمایش داده می شود.

الگوریتم ترکیبی COMRANK – ادامه

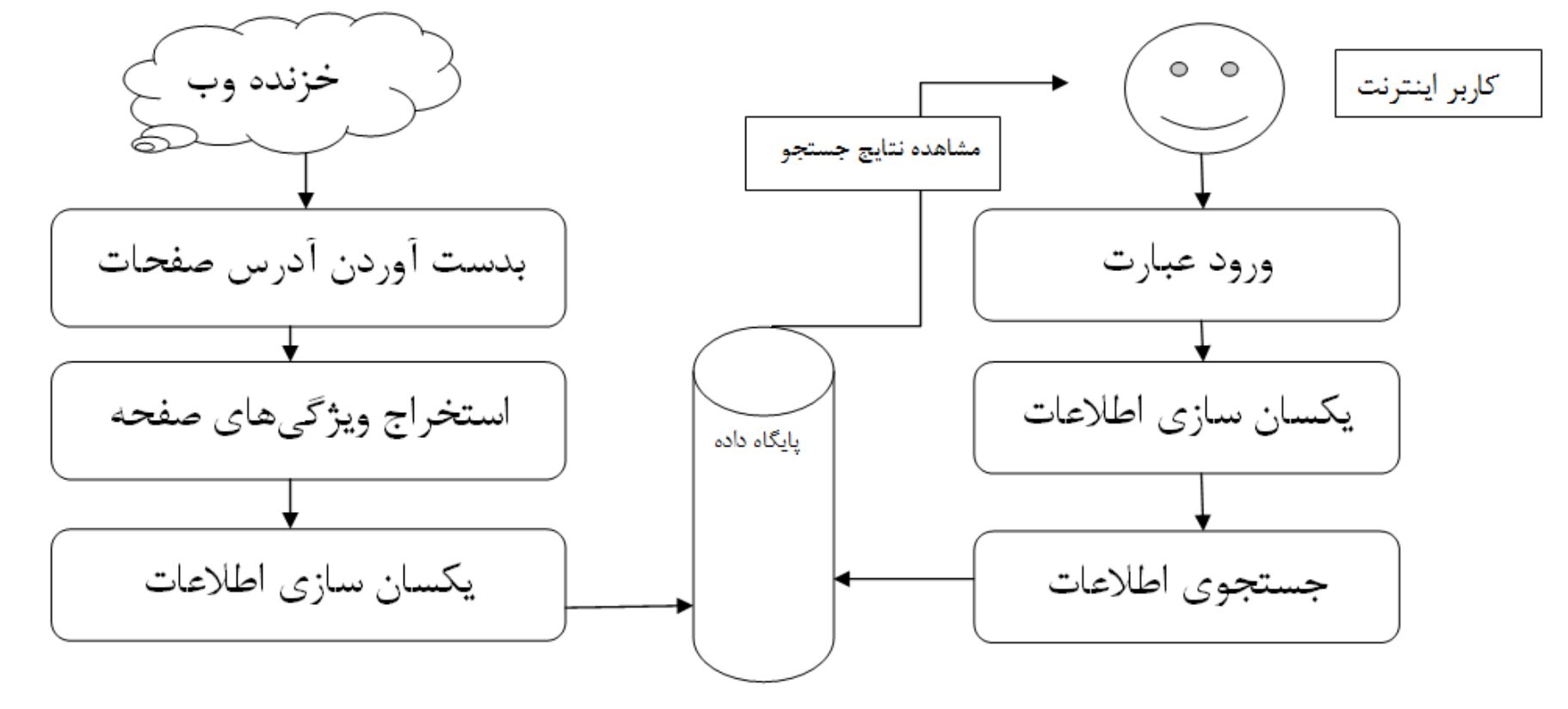


الگوریتم COMRANK

- رتبه بندی این الگوریتم بر اساس عملگر OWA صورت می پذیرد. هر یک از وزن ها باید دارای دو شرط زیر باشند:
  - مقدار هر یک از وزن ها در بازه بین ۰ و ۱ باشد
  - مجموع وزن ها برابر ۱ است.
- یکی از روش های یادگیری الگوریتم ComRank بوردا نام دارد و بر اساس خروجی مطلوب که از کاربر دریافت می کند وزن های خود را اصلاح می کند.
- در مجموع سرعت این روش نسبت بسیار کم است. به این دلیل که امتیاز باید برای تمام صفحات به ازا عبارت جستجو شده محاسبه شود و در صورت وجود تعداد صفحات زیاد، این الگوریتم کند عمل می کند.
- رتبه بندی در این الگوریتم بر اساس مجموع امتیازات هر صفحه و مطابق فرمول زیر است:

$$Rank(p) = \sum_{i=1}^{n} W_i R_i$$

پیاده سازی موتورجستجوی فارسی – معماری



مشکلات نگارشی زبان فارسی

- تنوع نحوه استفاده از" می "چسبان و غیر چسبان ، مثل کلمات" می تواند "و" میتواند."
- تنوع نحوه بکاربردن چسبان و غیر چسبان" ها "، مثل" آن ها "و" آنها."
- بکار بردن بعضی پیشوندها و پسوندها، مثل" همین که " و" همینکه "و یا " هیچ یک "و" هیچیک "و یا " راه گشا "و "راهگشا."
- بکاربردن همزه بصورت های مختلف، مثل" مسؤول "و" مسئول "یا" مسأله "و "مسئله."
- استفاده یا عدم استفاده از" ء "، برای کلمات مختوم به های بیان حرکت، در حالت مضاف، مثل" خانهٔ مسکونی" و "خانه مسکونی."
- تنوع استفاده از" ی "در کلمات عربی مختوم به" ا "، مثل" موسی "و" موسا."
- تنوع املایی بعضی کلمات که همه درست هستند، مثل" اتاق "و" اطاق."

مشکلات زبان فارسی – ادامه

- استفاده از کلمات اروپایی بصورت زبان اصلی یا ترجمه فارسی بخصوص در متون علمی، مثل "بروزرسانی" و" بروزآوری" برای کلمه update .
- استفاده یا عدم استفاده از جمع مکسر برای بعضی کلمات.
- تبدیل کلمات اروپایی به رسم الخط فارسی با همان تلفظ اصلی ، مثل سورس و source
- استفاده از" ا "و" آ "بجای هم ، مثل" فرایند "و" فرآیند."
- استفاده یا عدم استفاده از اعراب برای کلمات
- وجود چند کد کاراکتر برای یک حرف مانند حروف "ی" و "ک"
- استفاده یا عدم استفاده از نیم فاصله در کلمات
- وجود نوشته های محاوره ای در برخی از سایت ها و انجمن ها
- نوشتن کلمات فارسی با حروف انگلیسی (پینگلیش)

پیاده سازی خزنده وب فارسی – مشکلات

- وجود یا عدم وجود www در آدرس پیوندها
- دامنه های تکراری (اتصال چند دامنه به یک سایت) و اطلاعات کپی شده در سایت های مختلف
- صفحات با آدرس های داینامیک غیر موثر که باعث ایجاد صفحات تکراری می شوند.
  - پارامترهای پویا :
    - www.site.ir/index.php?id=10
  - چندآدرس برای یک صفحه:
    - www.site.ir/index.php?id=10&text=عبارت اضافه

پیاده سازی خزنده وب فارسی – مشکلات

- تفکیک پیوندها با فرمت های مختلف:
  - جاوااسکریپت
  - تصویر
  - فایل
  - بوکمارک
  - ایمیل
- وجود آدرس های نسبی در صفحات و تبدیل آنها به آدرس مطلق:
- page 1
- باید تبدیل شود به :
- page 1

پیاده سازی خزنده وب فارسی - ویژگیهای مناسب

- ویژگیهای یک خزنده خوب عبارتند از:
  - ایندکس نکردن صفحات تکراری
  - امکان اجرا در thread های موازی بسیار
  - گیر نیفتادن در چاله سایت ها:
    - از نظر سرعت
    - صفحات غیر مفید و اسپم
  - انطباق نرخ ایندکس کردن:
    - نرخ تغییرات سایت
    - سرعت سایت

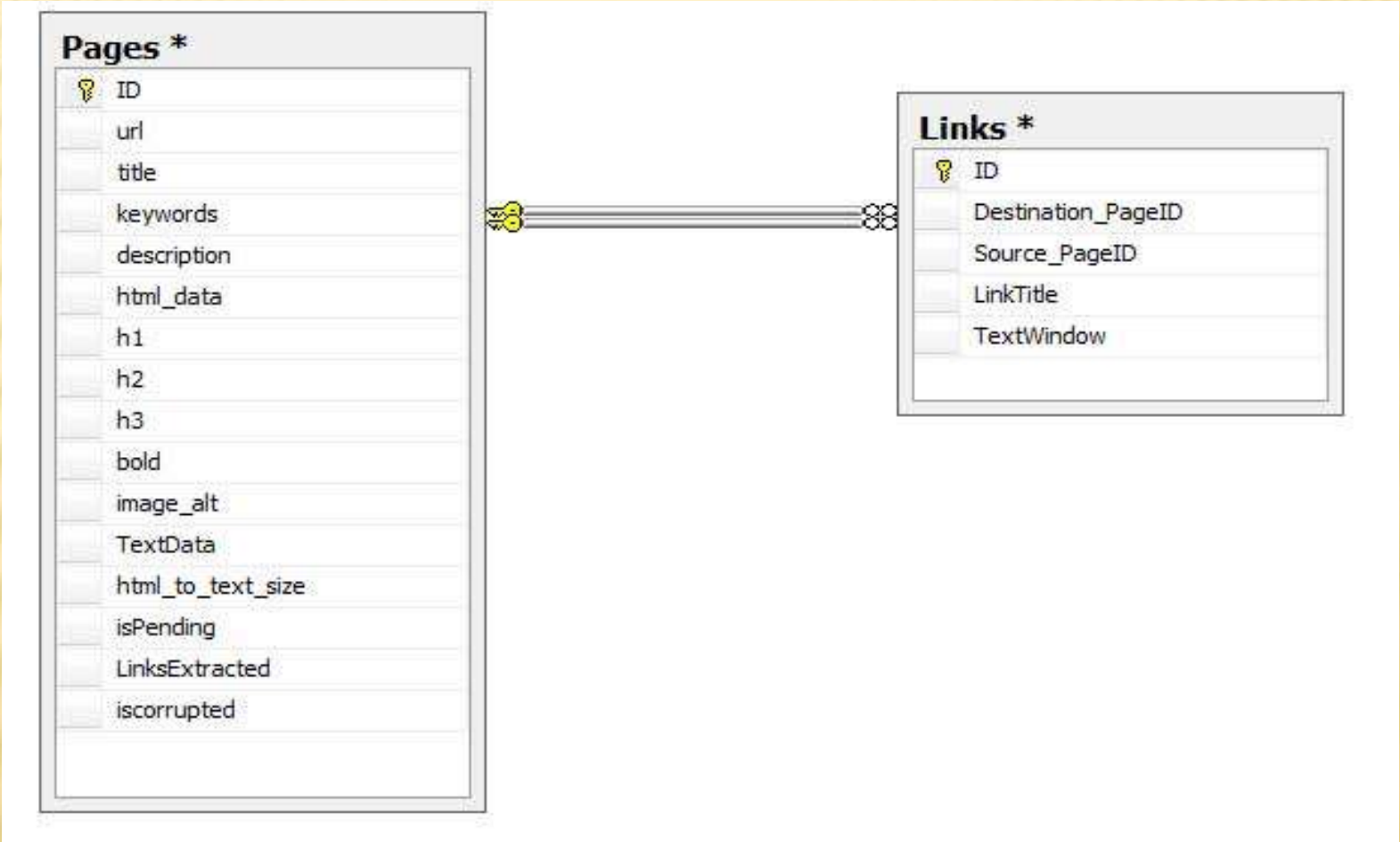
پیاده سازی خزنده وب فارسی

- در ابتدا ۲۰ سایت به عنوان مجموعه دانه به خزنده داده شد.
- خزنده فارسی در مدت هفت شبانه روز در حدود ۶۰۰.۰۰۰ صفحه را استخراج کرد.
- خزنده با استفاده از عبارات منطقی پیوندهای مفید صفحات را استخراج کرده و در بانک اطلاعاتی ذخیره کرد. اطلاعات استخراج شده در جداول مربوط به صفحه و پیوندها ذخیره می شوند.
- پس از عملیات جمع آوری اطلاعات، نرم افزار ایندکسر اطلاعات مفید را از صفحات استخراج و در فیلدهای اطلاعاتی مربوطه ذخیره کرد.

پارامترهای جستجوی فارسی وب

- پارامترها مبتنی بر پیوند:
  - عنوان پیوند (در متن و تصویر)
  - پنجره اطراف پیوند (به طول ۲۵۰ کاراکتر)
  - زمان بوجود آمدن پیوند (استفاده نشد)
- پارامترهای مبتنی بر متن:
  - عنوان صفحه
  - اطلاعات تگ های H1 تا H6
  - اطلاعات تگ B
  - توضیحات تصاویر
  - تگ متا (توضیحات و کلیدواژه ها)
  - چگالی HTML به متن صفحه
  - آدرس صفحه

ساختار جداول صفحات و پیوندها



شناسایی صفحات فریب آمیز

- برای شناسایی صفحات فریب آمیز از سه پارامتر استفاده شد:

1. نسبت طول HTML صفحات به متن آنها:

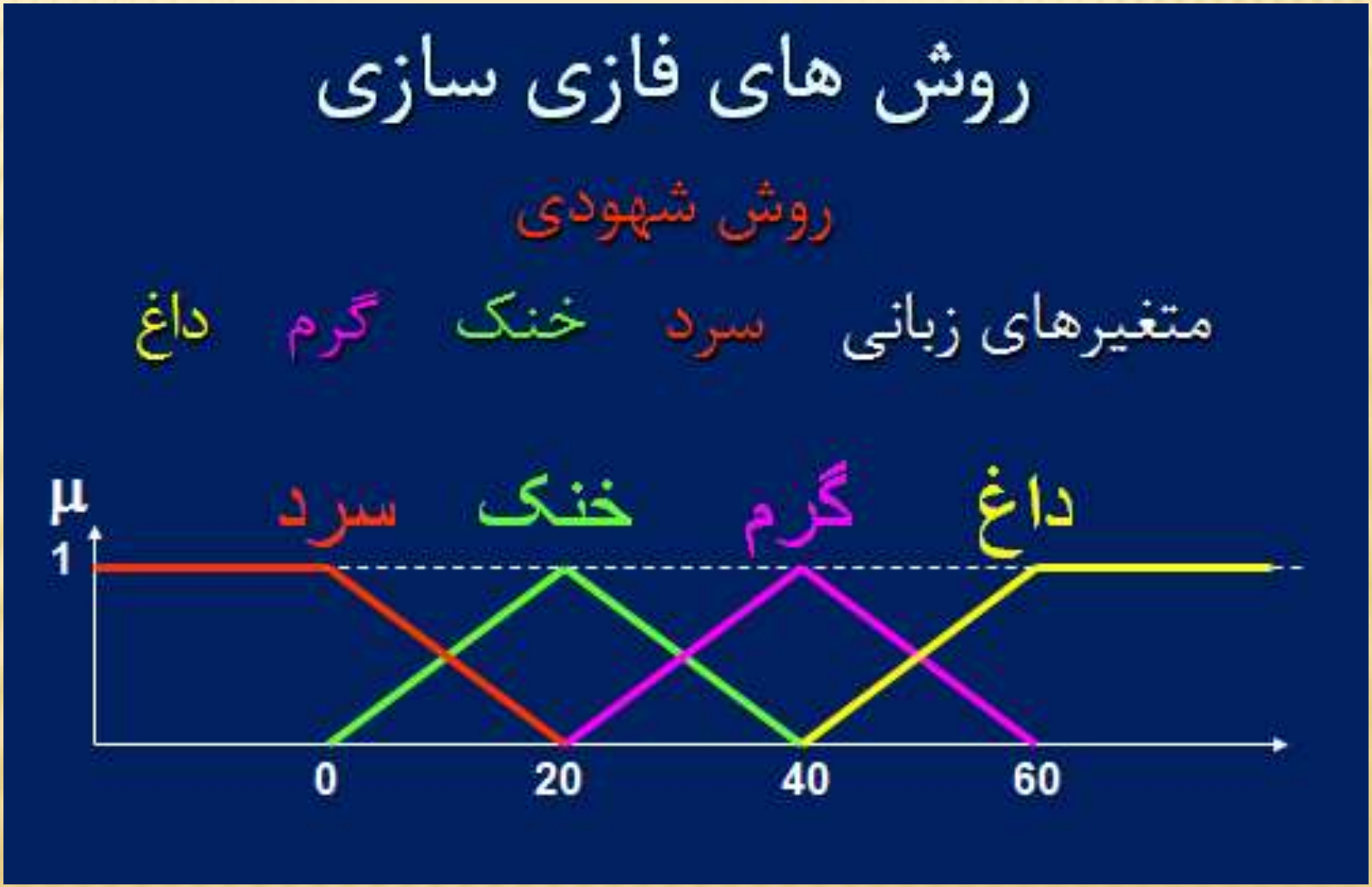
هر چه این میزان بیشتر باشد، صفحه اطلاعات کمتری دارد و احتمال اینکه یک صفحه فریب آمیز باشد بیشتر است.

2. نسبت حجم صفحه به حجم صفحه بدون در نظر گرفتن کلمه جستجو شده:

هر چه این میزان بیشتر باشد نشان دهنده این است که کلمه مورد نظر بیش از حد معمول در صفحه تکرار شده است و احتمال فریب آمیز بودن صفحه افزایش پیدا می کند. به صورت تجربی حد تشخیص فریب آمیز بودن صفحه برابر ۱.۰۴ در نظر گرفته شد.

3. استفاده از سیستم خبره فازی برای بدست آوردن میزان غیر فریب آمیز بودن صفحات.

سیستم خبره فازی – روش شهودی فازی سازی



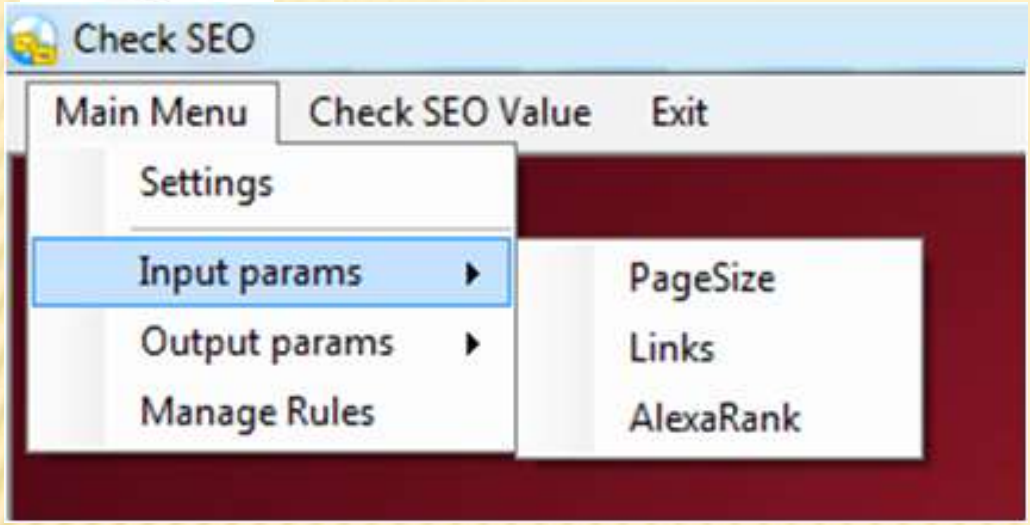
سیستم خبره فازی – تعاریف

اجزای سیستم خبره فازی

برخی از روش های فازی سازی:

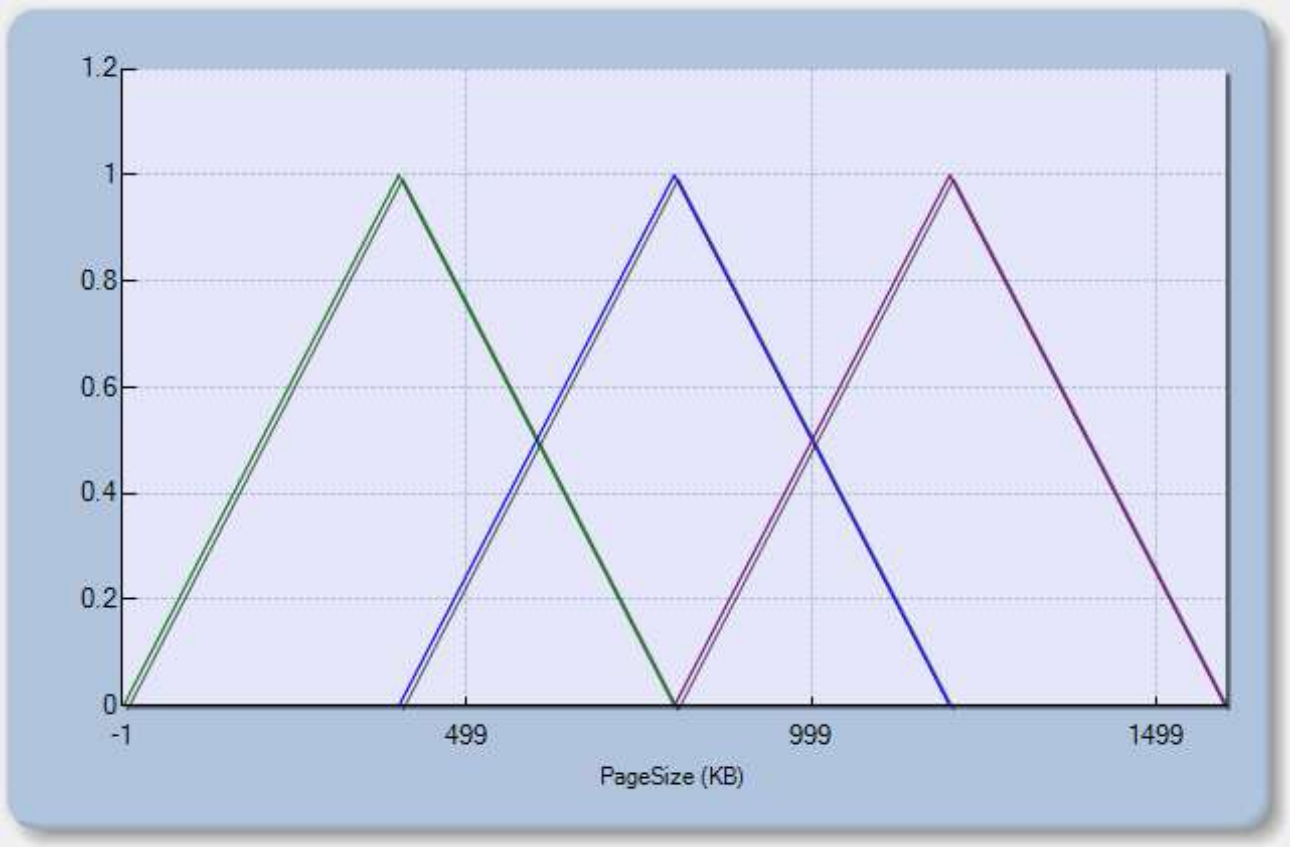
- روش شهودی
- استنتاج
- مرتب سازی آماری
- شبکه های عصبی
- الگوریتم های ژنتیک

سیستم خبره فازی جهت بدست آوردن میزان غیر فریب آمیز بودن صفحات



- تصمیم گیری بر اساس منطق فازی
- پارامترهای استفاده شده:
  - حجم اطلاعات صفحه
  - تعداد لینک های ورودی
  - رتبه سایت در الکسا
- تبدیل ورودی ها به اعداد فازی
- خروجی سیستم خبره احتمال غیر فریب آمیز بودن صفحه است.
- ورودی ها می توانند اعداد فازی به صورت مثلثی، ذورنقه ای و گووسی باشند.
- تصمیم گیری فازی بر اساس قوانین تعریف شده در سیستم صورت می گیرد، محدوده تعاریف اعداد فازی در قوانین اعمال می شود.
- توابع اجتماع، اشتراک، ترکیب فازی، برش لامدا و غیرفازی سازی در این سیستم قابل تنظیم است و می توان بر اساس نتایج بهترین حالت را تخمین زد.
- خروجی این سیستم می تواند به عنوان یک پارامتر در رتبه بندی صفحات استفاده شود.

تعریف محدوده اعداد فازی برای حجم صفحات



سبز : کم

آبی : متوسط

قرمز: زیاد

تعریف قوانین سیستم خبره فازی

Page Size	Links	Alexa	SEO Rank
Low	Low	Low	Medium
Low	Medium	Low	Medium
Low	High	Low	High
Low	Low	Medium	Medium
Low	Low	High	High
Low	Medium	Medium	Medium
Low	Medium	High	Medium
Low	High	High	High
Low	High	Medium	High
Medium	Low	Low	Low
Medium	Medium	Low	Medium
Medium	High	Low	High
Medium	Low	Medium	Medium
Medium	Low	High	Low
Medium	Medium	Medium	Medium
Medium	Medium	High	Medium
Medium	High	Medium	High
Medium	High	High	Medium
High	Low	Low	Medium
High	Low	Medium	Low
High	Low	High	Low
High	Medium	Low	Low
High	Medium	Medium	Medium
High	Medium	High	Low
High	High	Low	High
High	High	Medium	Medium
High	High	High	Low

پیاده سازی موتور جستجوی وب فارسی

- پیاده سازی موتورجستجو در دو مرحله انجام شد:
  - یادگیری و محاسبه وزن های پارامترهای مختلف:
    - استفاده از افراد خبره
    - شبکه های عصبی
    - الگوریتم های ژنتیک
    - الگوریتم یادگیری بوردا
  - پیاده سازی تابع رتبه بندی

پیاده سازی موتور جستجوی وب فارسی

- برای پیاده سازی موتور جستجو از کتابخانه Microsoft Full Text Search استفاده شد.
- این کتابخانه نسبت به توابع SQL سرعت بسیار بیشتری دارد و امکان جستجوی عبارات مشابه، جستجوی ترکیبی کلمات، بدست آوردن وزن عبارت و رتبه بندی صفحات را مهیا می سازد.
- رتبه بندی صفحات بر اساس حاصل ضرب امتیاز صفحه برای هر پارامتر و وزن پارامتر مورد نظر صورت می گیرد:

$$Rank(p) = \sum_{i=1}^{n} W_i R_i$$

نمونه ای از وزن پارامترها

- پس از انجام آزمایشات مختلف بر روی داده ها، وزن های مناسب برای پارامترها به صورت زیر در نظر گرفته شد:
  - Ánchor: 0.4
  - Title: 0.1
  - Meta Keywords: 0.1
  - Meta description: 0.1
  - H1,H2,H3 : 0.1
  - Bold: 0.05
  - Image Alt attribute: 0.05
  - Page Data: 0.1
  - HTML / Text (Spam Penalty) : -0.1
  - Search term density (Spam Penalty) : -1
  - SEO Rank (Spam Penalty) : 0
  - Search term exist in URL : 0

پیاده سازی موتور جستجوی وب فارسی – اندازه گیری خطا

- روش های اندازه گیری خطا:
- بدست آوردن فاصله تا نتیجه مطلوب اول
- در نظر گرفتن تعداد نتایج مشترک در ۱۰ نتیجه اول جستجو و همچنین فاصله هر یک از نتایج با نتیجه مطلوب

پیاده سازی موتور جستجوی وب فارسی - اندازه گیری خطا

- برای اندازه گیری خطا از داده های WEBIR و پرس و جوهای ارائه شده در آن استفاده شد.
- در داده های WEBIR پرس و جوهای آزمایشی به همراه جواب مناسب داده شده است. می توان پاسخ نهایی را با آنها مقایسه کرده و مقدار خطا را محاسبه کرد.
- برای اندازه گیری خطا از ۱۰ پرس و جوی نمونه استفاده گردید که نتایج مطلوب آنها در داده های WEBIR مشخص شده است.
- اندازه گیری خطا بر اساس فاصله ۳ نتیجه اول داده های WEBIR با نتایج خروجی انجام شد.

اندازه گیری خطا – داده های WEBIR

- مشکلات داده های WEBIR:
  - داده های WEBIR قدیمی است.
  - برخی از فیلدهای مفید در اطلاعات وجود ندارد. به عنوان مثال پارامتر بسیار مهم Anchor . لذا برای استفاده از این داده ها، پارامترها مجددا استخراج شدند.
  - پاسخ های مناسب بر اساس ترکیب الگوریتم های مختلف بدست آمده اند و نظارت انسانی روی آنها نبوده لذا پاسخ های مناسب داده های فوق لزوما بهترین پاسخ مورد نظر نیستند.

نتیجه گیری

- با بررسی پارامترهای مختلف مبتنی بر متن و مبتنی بر پیوند به این نتیجه رسیدیم که در حالتی که وزن پارامترهای مبتنی بر پیوند بیشتر باشد سیستم پاسخ مناسب تری تولید خواهد کرد. البته این موضوع منوط به داشتن داده های مناسب است.
- با وجود اینکه الگوریتم comRank در آزمایشات انجام شده نتایج بهتری نسبت به روش های سنتی دارد اما در عمل نیازمند محاسبات بیشتری است و باعث کندی در پاسخ دادن به کاربر می شود.
- در صورتی که موتور جستجو بتواند نتایج خود را با خواسته های کاربر منطبق کند رضایتمندی بیشتری در کاربر ایجاد خواهد کرد. لذا پاسخ جستجو می تواند برای کاربران مختلف متفاوت باشد.

مراجع و منابع

[1] http://nlp.stanford.edu/IR-book/html/htmledition/the-web-graph-1.html

[2] http://www-personal.ksu.edu/~eddery/Google.pdf

[3] J. A. Aslam, and M. Montague, "Models forMetasearch". In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New Orleans, USA, September 2001, 2001, pp. 276-284.

[4] B. T. Bartell, "Optimizing Ranking Functions: A Connectionist Approach to Adaptive Information Retrieval",PhD Theis, University of California, San Diego, USA, 1994.

[5] McBryan, O.A. "GENVL and WWWW: Tools for Taming the Web",In Proceedings of the First International World Wide Web Conference, pp. 79-90,1994.

[6] Gulli, A., Signorini , A., "The Indexable web is more than 11.5 billion pages", In Proceedings of the 14th international conference on World Wide Web, pp. 902-903, ACM Press, 2005.

[7] http://www.internetnews.com/stats/article.php/1363881

[8] http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf

[9] Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern Information Retrieval. ACM Press/ Addison-Wesley.

[10] Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. Information Processing and Management, 24(5), 513–523.

مراجع و منابع – ادامه

[11] Robertson, S. E., Walker, S., HancockBeaulieu, M. M., Gatford, M., & Payne, A. (1995). Okapi at TREC-4. In NIST Special Publication. The Fourth Text Retrieval Conference (TREC-4) (pp. 73—96).

[12] Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. Technical report, Computer Science Department,
Stanford University.

[13] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM,46(5),604–632.

[14] Zareh Bidoki, A. M., & Yazdani, N. , DistanceRank: An intelligent ranking algorithm for web pages, Information Processing and Management (2007), doi:10.1016/j.ipm.2007.06.004.

[15] Najork, M., Zaragoza, H., & Taylor, M. J. (2007). Hits on the web: how does it compare? In Proceedings of SIGIR'07 (pp. 471-478).

[16] Yager, R.R. (1988). On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Transactions on Systems, Man and Cybernetics,18(1).

[17] Filev D. and Yager R. R., "On the issue of obtaining OWA operator weights", Journal of Fuzzy Sets and Systems, Vol. 94, No. 2, pp. 157-169, 1998.

[18] Cho, J., Roy, S., & E. Adams, R. (2005). Page Quality: In Search of an Unbiased Web Ranking. In Proceedings of ACM International Conference on Management of Data (SIGMOD).

[19] Filev, D., and Yager, R. R., 1994. Learning OWA operator weights from data. Proceedings. of the third IEEE Conference on Fuzzy Systems, Vol.1, pp. 468-473.

[20] LETOR, http://research.microsoft.com/users/tyliu/LETOR/default.aspx

با تشکر و سپاس از حضور و توجه شما

رضا شیرازی مفرد

www.rezashirazi.com