ہمکنار پاکستان

Urdu Search Engine

This is a technical report of comparing other language based search engines with our search engine in terms of design and implementation

High Performance Computing and Networking Lab
Center for Language Engineering
Al-Khawarizmi Institute Of Computer Science,
University Of Engineering and Technology, Lahore

Contents

1 Introduction

This is a technical report of comparing popular search engines with our search engine in terms of design, implementation and features. In order to compare our proposed search engine with similar ones, we put together a list of search engines on the basis of their popularity and language features. As there is no Urdu language based search engine available so we have selected Persian and Arabic search engines for language comparison because these languages are similar to Urdu. The search engines we have selected for comparison are Google, Yahoo, Baidu, DuckDuckGo, Common Search, Yooz, Rismoon, Salam, Parsijoo, Yamli, Yoolki and Eiktub. We are comparing the technologies they are using, how they are computing their big data so fast, where they are storing their data, how they are crawling the web, how they are indexing the crawl data, how they are showing results on the basis of ranking, how they are managing such a large traffic and what kind of language features they are providing. Some search engines also provide advertising services from which website publishers can earn money by displaying advertisement on their websites.

2 Search Engines

Search engine is a program which return search results (documents) for a specific search term or a query. Typically, search engine uses a program crawler to fetch as many documents as possible from the web. Than another program indexer is used to read these fetched documents and create an index based on the fetched documents word just like an index of a book. Than each search engine algorithms use this index to search and return searched documents. They also create or use algorithms to return only meaningful results like for ranking or trending results.

Search engine history all started in 1990 with Archie, an FTP site hosting an index of downloadable directory listings. Search engines continued to be primitive directory listings, until search engines developed to crawling and indexing websites, eventually creating algorithms to optimize relevancy.

Yahoo started off as just a list of favorite websites, eventually growing large enough to become a searchable index directory. They actually had their search services outsourced until 2002, when they started to really work on their search engine [1].

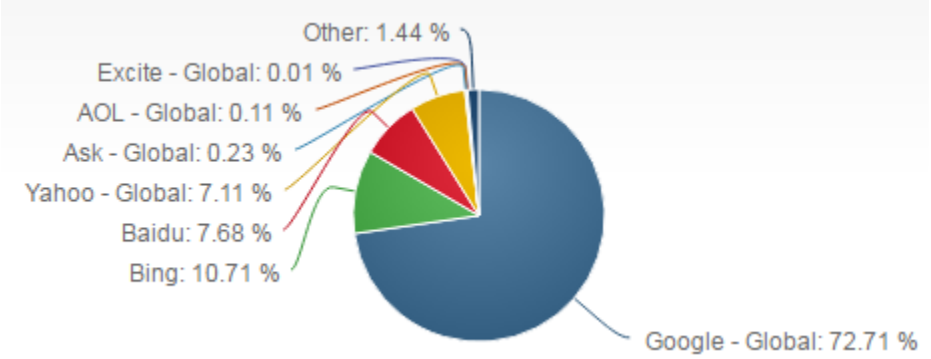

Fig 1: Global Market Share from January, 2016 to February, 2017

From figure1, we can see that Google has the largest market share globally from January, 2016 till February, 2017 and is equal to 72%. Bing is at the second position with 10.71% of the market share. Baidu and Yahoo are close and are on third position with 7.68% and 7.11% respectively. Then comes Ask,AOL and Excite with 0.23%,0.11% and 0.01% of Market share, respectively. 1.44% of Market share are in possession with some small multiple companies.

3 Top Search engines design and implementation

In this section, we look at some globally top search engines.

3.1 Yahoo:

Yahoo! Search is a web search engine owned by Yahoo. As of February 2015, it is the third largest search engine in the US by the query volume at 12.8%, after its competitors Google at 64.5% and Bing at 19.8%.

Originally, "Yahoo Search" referred to a Yahoo-provided interface that sent queries to a searchable index of pages supplemented with its directory of websites. The results were presented to the user under the Yahoo! brand. Originally, none of the actual web crawling and data housing was done by Yahoo! itself. In 2001, the searchable index was powered by Ink Tomi and later was powered by Google until 2004, when Yahoo! Search became independent. On July 29, 2009, Microsoft and Yahoo! announced a deal in which Bing would henceforth power Yahoo! Search [2].

In backend, Yahoo's infrastructure harnesses Hadoop Distributed File System (HDFS) for ultra-scalable storage, Hadoop MapReduce for massive ad-hoc batch processing, Hive and Pig for database-style analytics, HBase for key-value storage, Storm for stream processing, and Zookeeper for reliable coordination. For crawling, they use their own crawler Yahoo Slurp and sometimes Bing Bot [3] [4].

For frontend, they use The Yahoo! User Interface Library (YUI) which is an open-source JavaScript library for building richly interactive web applications using techniques such as Ajax, DHTML, and DOM scripting. YUI includes several core CSS resources. It is available under a BSD License [5].

Yahoo Ranking Algorithm:

Yahoo Algorithm is not far from Google Algorithm but different at some points, Yahoo gives much interests in taking its web directory as part of the its Ranking Algorithm. It is sad but real, if you are not in Yahoo Web Directory, you can wait a long time before you appear in Yahoo SERPS.

If you have a business website, you will have to pay $299 dollars annually to remain inside Yahoo Web Directory. For organizations or informational web sites you can try to Submit freely your web site to Yahoo Web Directory.

Although it's not by applying a fast submission to Yahoo Web Directory and after forgetting about it that your web site will rank in the TOP 10 for each keyword you target! You will have to do some follow up with the Yahoo editors in order to get your website inside!

Yahoo Ranking Rules:

- The title of your website must contain your major keywords. The title is the biggest ranking factor of Yahoo's raking algorithm!
- The description you give when submitting to yahoo web directory should include your major keywords, however don't try to repeat them too much. Very important step.
- Click Popularity is part of Yahoo's Algorithm, Google doesn't put much on that to determines one website ranking. The more visitors click on your website from Yahoo SERP's, the more you'll get close to the Top Ranking.
- The category you are listed in Yahoo Web Directory should (if possible) contain some keywords. This plays a little with the ranking.
- Site-wide linking. It’s a choice you have to make, you can get one link per domain in order to rank well on google search engine but you can also get a lot of site-wide linking from another site, Yahoo loves it [6].

3.2 Google:

Google Search, commonly referred to as Google Web Search or simply Google, is a web search engine developed by Google. It is the most-used search engine on the World Wide Web, handling more than three billion searches each day. As of February 2016, it is the most used search engine in the US with 64.0% market share [7].

Old Backend Design:

More than a decade ago, Google built a new foundation for its search engine. It was called the Google File System GFS, for short and it ran across a sweeping army of computer servers, turning an entire data center into something that behaved a lot like a single machine.

After Google released research papers describing GFS and a sister software platform called MapReduce the piece that crunches the data Yahoo, Facebook, and others built their own version of the Google foundation. It was called Hadoop, and this open source platform is now driving a revolution across the world of business software as well [8]. Figure 2 shows Google old design vs Open Source platform [9]. As you can see from figure Google File System (GFS) is equal to Hadoop, Big table is equal to Hbase and Google Mapreduce equal to Hadoop Mapreduce

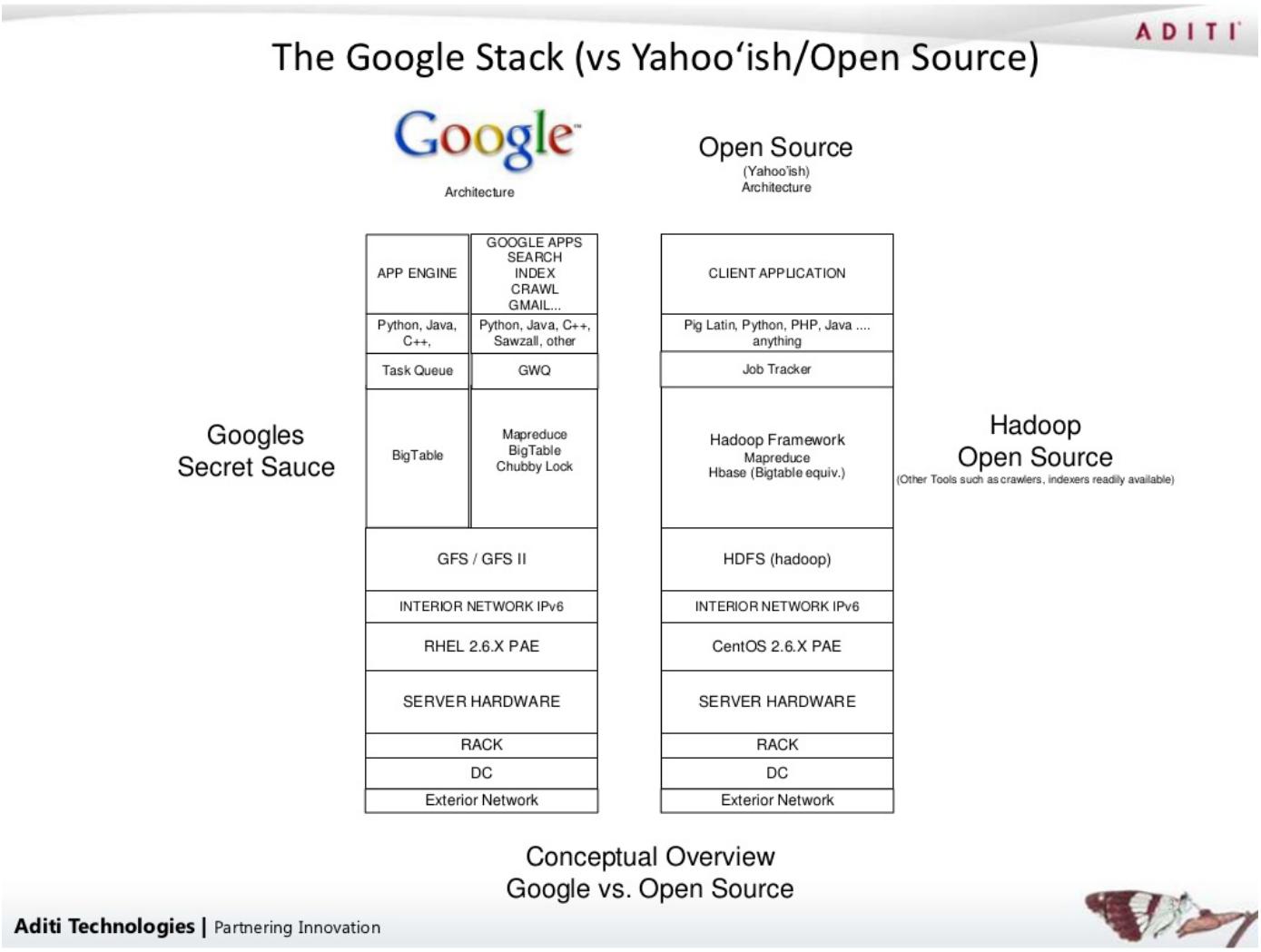

Figure 2: Google old design vs Open Source Platform

New Backend Design:

But now Google no longer uses GFS. The company moved its search to a new software foundation based on a revamped file system known as **Colossus**. Colossus now underpins virtually all of Google's web services, from Gmail, Google Docs, and YouTube to the Google Cloud Storage service the company offers to third-party developers. Whereas GFS was built for batch operations i.e., operations that happen in the background before they're actually applied to a live website Colossus is specifically built for "real-time" services, where the processing happens almost instantly.

With its search engine, Google has not only dropped GFS. It has dropped MapReduce. Rather than using MapReduce to build a new index every so often, it uses a new platform called "**Caffeine**" that operates more like a database, where you can read and write data whenever you like. It is a new web

indexing system. Caffeine provides 50 percent fresher results for web searches than our last index, and it's the largest collection of web content we've offered. Whether it's a news story, a blog or a forum post, you can now find links to relevant content much sooner after it is published than was possible ever before.

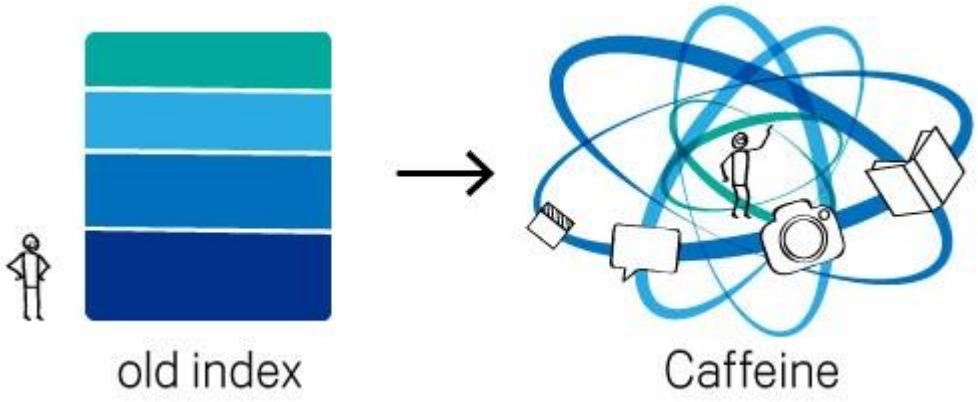

Figure 3: old index vs new index(Caffeine)

Their old index had several layers, some of which were refreshed at a faster rate than others. The main layer would update every couple of weeks. To refresh a layer of the old index, they would analyze the entire web, which meant there was a significant delay between when they found a page and made it available to you.

With Caffeine, they analyze the web in small portions and update their search index on a continuous basis, globally. As they find new pages, or new information on existing pages, they can add these straight to the index. That means we can find fresher information than ever before no matter when or where it was published.

Caffeine lets them index web pages on an enormous scale. In fact, every second Caffeine processes hundreds of thousands of pages in parallel.

They have built Caffeine with the future in mind. Not only is it fresher, it's a robust foundation that makes it possible for them to build even faster and comprehensive search engine that scales with the growth of information online, and delivers even more relevant search results to us [8][10].

The Google Stack:

In figure 4, “What does it take to make Google work at scale?” Schwarzkopf discusses the architecture behind those 139 microseconds between submitting a search request in the Google input bar, and the pages of search results that are returned.

All of what happens, takes place in containers between customized Linux kernels on each data machine and the transparent layer of distributed systems.

He identifies 16 different software technologies that work in tandem to return the real-time, contextual, personalized search results that users expect from Google [11] [12].

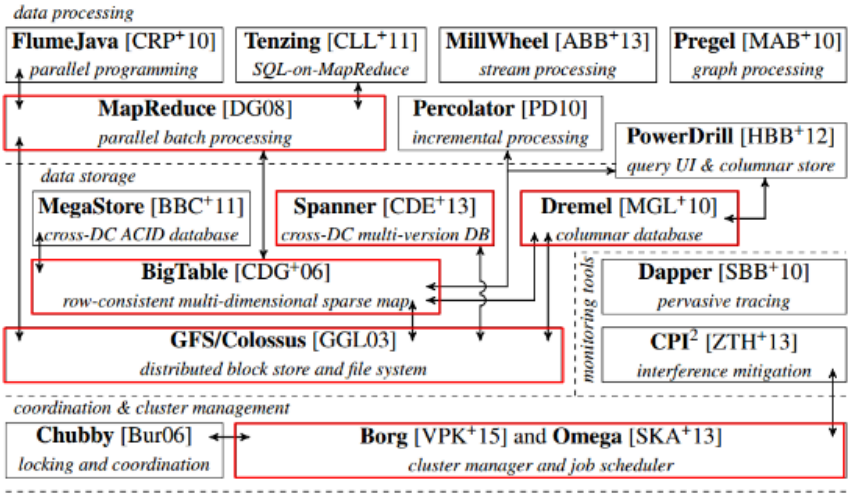

Figure 4: The Google Stack

Google Stack include these:

- **GFS/Colossus:** a bulk block data storage system.
- **Big Table:** a three-dimensional key-value store that combines row and column keys with a timestamp.
- **Spanner:** Software that uses the GPS and atomic clocks within data centers to enable transactional consistency at a global scale.
- **MapReduce:** a parallel programming framework.
- **Dremel:** a column-oriented data store useful for quick, interactive queries.
- **Borg/Omega:** the father of Kubernetes, a cluster manager and scheduler for large-scale, distributed data center architecture.

Software Define Network (SDN):

Obviously, Google operates a massive network. It's so large that a 2010 study by Arbor Networks concluded, "If Google were an ISP, it would be the fastest growing and third largest global

carrier. Only two other providers (both of whom carry significant volumes of Google transit) contribute more inter-domain traffic."

Google uses a combination of Quagga open source software (**Quagga** is a network routing software suite providing implementations of Open Shortest Path First (OSPF), Routing Information Protocol (RIP), Border Gateway Protocol (BGP) and IS-IS for Unix-like platforms, particularly Linux, Solaris, FreeBSD and NetBSD) along with OpenFlow to optimize its data center interconnects. Google calls its SDN network "**B4**."

OpenFlow is a technique for controlling network operations in software run on centralized computer servers saving cost, time and power. It aims to simplify and virtualize today's business networks that currently require a number of specialized, distributed systems, each with its own software load.

The growth of Google's back-end (east-west) network is quickly surpassing its front-end user-facing network. This growth is expensive because the network doesn't scale economically like storage and compute do. The operating expense of compute and storage becomes cheaper per unit as scale increases, but this is not the case with the network.

Google's rationale for software-defined networking are. First, by separating hardware from software, the company can choose hardware based on required features while being able to innovate and deploy on software timelines. Second, it provides logically centralized control that will be more deterministic, more efficient and more fault-tolerant. Third, automation allows Google to separate monitoring, management and operation from individual boxes. All of these elements provide flexibility and an environment for innovation [13].

Google page rank:

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites [14].

Suppose a small universe of four web pages: **A**, **B**, **C** and **D**. If all those pages link to **A**, then the **PR** (PageRank) of page **A** would be the sum of the **PR** of pages **B**, **C** and **D**.

$PR(A) = PR(B) + PR(C) + PR(D)$

But then suppose page **B** also has a link to page **C**, and page **D** has links to all three pages. One cannot vote twice, and for that reason it is considered that page **B** has given half a vote to each. In the same logic, only one third of **D**'s vote is counted for A's PageRank.

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}$$

In other words, divide the **PR** by the total number of links that come from the page.

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}$$

Finally, all of this is reduced by a certain percentage by multiplying it by a factor **q**. For reasons explained below, no page can have a PageRank of 0. As such, Google performs a mathematical operation and gives everyone a minimum of 1 - *q*. It means that if you reduced 15% everyone you give them back 0.15.

$$PR(A) = \left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots \right) q + 1 - q$$

So, one page's PageRank is calculated by the PageRank of other pages. Google is always recalculating the PageRanks. If you give all pages a PageRank of any number (except 0) and constantly recalculate everything, all PageRanks will change and tend to stabilize at some point. It is at this point where the PageRank is used by the search engine [30].

Google search Ranking:

Google has multiple named parts of the algorithm that influence search rankings. Some of them are:

1. **Google Panda:**

Launched on 2011. Google Panda takes the quality of a site's content into account when ranking sites in the search results. For sites that have lower quality content, they would likely find themselves negatively impacted by Panda. As a result, this causes higher quality content to surface higher in the search results, meaning higher quality content is often rewarded with higher rankings, while low-quality content drops.

2. **Penguin:**

Launched on 2012. The second major Google algorithm is Penguin. Penguin deals solely with link quality and nothing else. Sites that have purchased links or have acquired low-quality links through places such as low-quality directories, blog spam, or link badges and infographics could find their sites no longer ranking for search terms.

3. **Humming Bird:**

Launched on 2013. Hummingbird is an algorithm to ensure they are serving the best results for specific queries. Hummingbird is more about being able to understand search queries better, particularly with the rise of conversational search.

It is believed that Hummingbird is positively impacting the types of sites that are providing high-quality content that reads well to the searcher and is providing answers to the question the searcher is asking, whether it is implied or not.

Hummingbird also impacts long-tailed search queries, similarly to how Rank Brain is also helping those types of queries. Google wants to ensure that they can provide high-quality results for the longer queries. For example, instead of sending a specific question related to a company to the company's

homepage, Google will try to serve an internal page on the site about that specific topic or issue instead [15].

4. Pigeon:

Launched on July 24, 2014 for U.S. The "Pigeon Update" is a new algorithm to provide more useful, relevant and accurate local search results that are tied more closely to traditional web search ranking signals. Google stated that this new algorithm improves their distance and location ranking parameters. The algorithm affects search results within both Google Maps Search & Google Web Search. Searching Google and searching Google Maps has often provided a very different set of results. This changed after the rollout of Pigeon. The algorithm connects web search and map search in a more cohesive way [32] [33]. Figure 5 shows the google (www.google.com.pk) results for search term "cafe" and figure 6 shows the google (us) results for search term "cafe" by using proxy server of US. We can see the difference that google.com.pk shows Pakistan café results also in SERP with other results and google (US) shows the US café results in SERP.

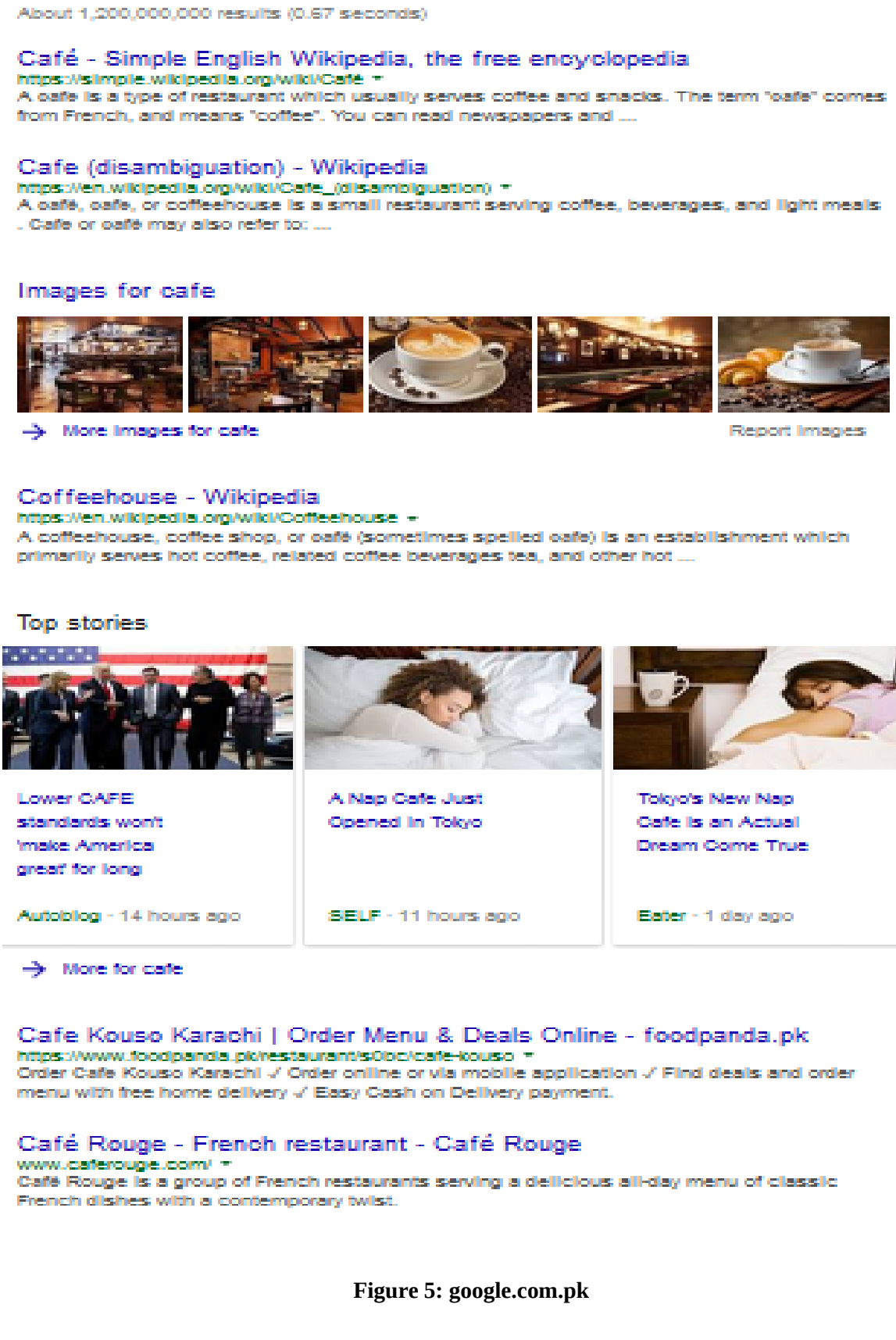

Figure 5: google.com.pk

Figure 6: Google (US)

5. **Rank Brain:**

Rank Brain was launched in early 2015 and is used globally by Google. Rank Brain is an artificial intelligence (AI) program used to help process Google search queries.

RankBrain uses artificial intelligence to embed vast amounts of written language into mathematical entities, called vectors, that the computer can understand.

If RankBrain sees a word or phrase it isn't familiar with, the machine can make a guess as to what words or phrases might have a similar meaning and filter the result accordingly, making it more effective at handling never-before-seen search queries.

Google told Search Engine Land that RankBrain favoured different results in Australia versus the United States for that query because the measurements in each country are different, despite the similar names [31].

Google Adsense:

Google uses its technology to serve advertisements based on website content, the user's geographical location, and other factors. AdSense has become one of the popular programs that specializes in creating and placing banner advertisements on a website or blog, because the advertisements are less intrusive and

the content of the advertisements is often relevant to the website. Many websites use AdSense to make revenue from their web content (website, online videos, online audio content, etc.), and it is the most popular advertising network. AdSense has been particularly important for delivering advertising revenue to small websites that do not have the resources for developing advertising sales programs and salespeople to seek out advertisers. To display contextually relevant advertisements on a website, webmasters place a brief Javascript code on the website's pages. Websites that are content-rich have been very successful with this advertising program, as noted in a number of publisher case studies on the AdSense website [29].

3.3 Common Search:

Common Search Engine is based on the ~150TB Common Crawl monthly dumps. Their main text index only needs to be refreshed every month, like Google did up until 2003.

The backend of Common Search contains all the Python code working on the raw data, analyzing it and indexing it into Elasticsearch.

They have 2 Elasticsearch clusters as you can see in figure 7, the largest one contains the main inverted index, mapping words to document IDs. The second one acts as a document store, mapping document IDs to their actual content.

User queries are sent to the frontend, a Go server that forwards these requests to the index and returns them properly formatted as HTML or JSON if JavaScript is supported by the client.

They update their index once a month, so that they can cache results aggressively in many edge locations around the world provided by a CDN, which lowers their costs and makes popular queries extremely fast. Their Elasticsearch cluster is using AWS CloudFormation and Backend is using the Spark EC2 scripts. CloudFormation allows them to provision an entire stack with instances and all their dependencies in just one command. They are currently using the Spark EC2 scripts to provision their Spark workers, though they aim to switch to CloudFormation in the future [16].

The frontend contains 2 main components:

- A Go server that receives user queries (as HTTP GETs for page loads or AJAX calls), sends them to an Elasticsearch index, and then returns results as HTML or JSON.
- Website is developed by using markdown language and pelican to generate the website.
- An optional JavaScript/CSS layer that provides a fast, single-page search experience to the otherwise static result pages.

Search Ranking:

The final Elasticsearch score of each document is currently the product of 2 factors: the query score (specific to the searched terms) and the static ranking score (computed at index time, and don't depend on

search terms. They are roughly equivalent to the "popularity" of each domain, though improving our document-level signals is a big priority).

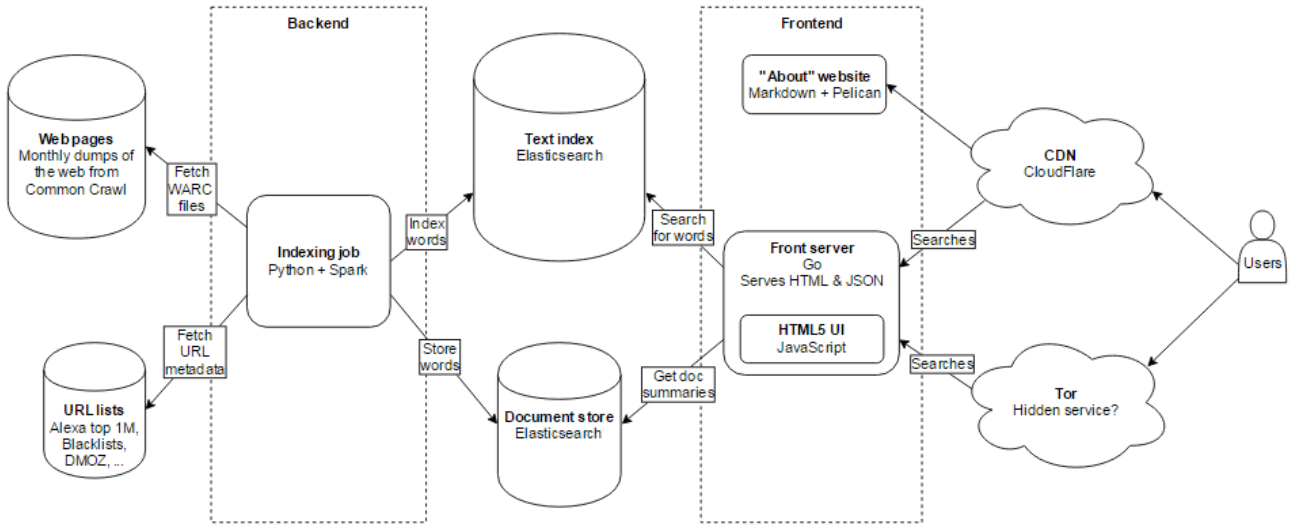

Figure 7: Common Search Architecture

3.4 Baidu:

Baidu is the number one search engine in China, with approximately 68% of the country's market share. Baidu also use Hadoop as a filesystem and HBase as a storage system like Yahoo. But Baidu also build Alluxio which is formerly known as Tachyon, the world's first memory-centric virtual distributed storage system. Baidu uses Alluxio to improve their data analytics performance by 30 times.

Currently within Baidu, they have a production of Tachyon cluster with 100 nodes and over 2PB of storage space – this cluster mainly serves as the cache layer for their big data analytics engine. Enabling interactive queries with Spark and Tachyon [17] [18].

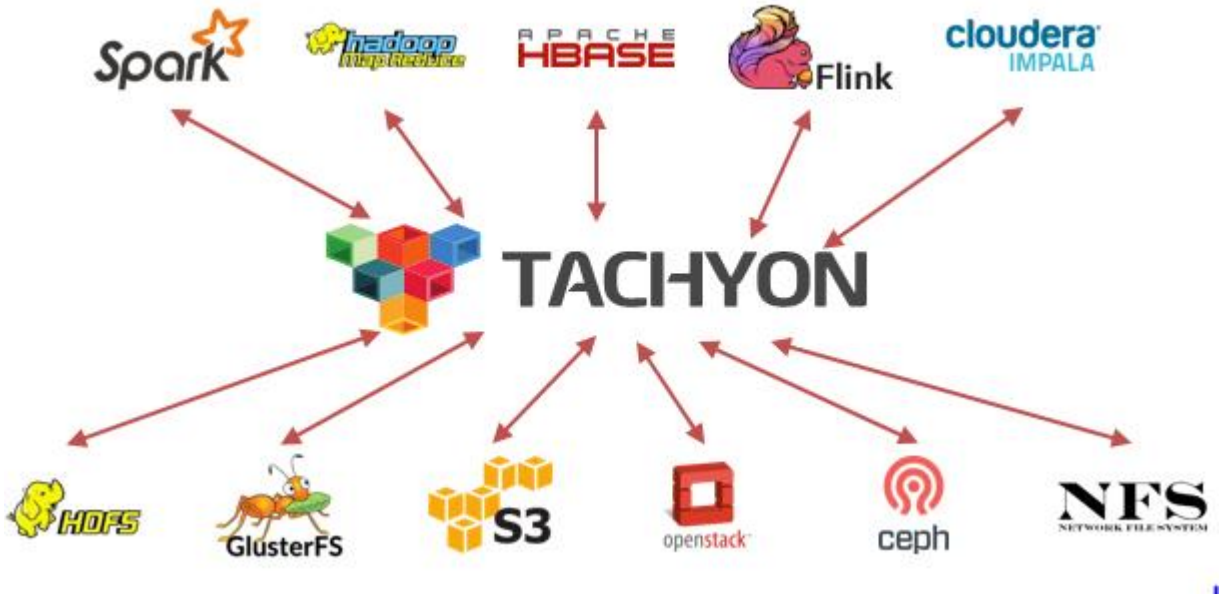

Figure 8: Tachyon

3.5 DuckDuckGo:

DuckDuckGo (DDG) is an Internet search engine that emphasizes protecting searchers' privacy and avoiding the filter bubble of personalized search results. DuckDuckGo distinguishes itself from other search engines by not profiling its users and by deliberately showing all users the same search results for a given search term. DuckDuckGo emphasizes getting information from the best sources rather than the most sources, generating its search results from key crowdsourced sites such as Wikipedia and from partnerships with other search engines like Yandex, Yahoo!, Bing, and Yummly [19].

The Platform:

- Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale cloud computing easier for developers)
- Perl & CPAN (repository archive)
- Server Density - monitoring
- Apache Solr
- PostgreSQL (for instant search: Wikipedia results)
- Memcached
- Bucardo - asynchronous PostgreSQL replication system
- Global Traffic Director - load balancing between regions

- Nginx
- getFavicon - serves favicons
- JavaScript
- YUI (moving to jQuery)
- Knockout (Web framework)

DDG is using EC2 for computation. Most components of DDG, including all front-end components, are now on AWS. They are moved to multiple regions for better front-end performance which make DDG fast everywhere and the users will come. One complaint users had is speed. By running DDG in 4 AWS datacenters (California, Virginia, Singapore, Ireland) DDG is able to be closer to its users around the world. Identical software runs in all datacenters.

Global Traffic Director is used for their DNS and to load balance users across regions. DDG would like to use more regions (South America and another Asian datacenter) but the Traffic Director currently works only in four regions.

Many databases are used for DDG, including PostgreSQL, Apache Solr, Berkeley, and flat files. Bucardo is used for PostgreSQL replication. Solr for indexing of documents. PostgreSQL masters are in the basement and the slaves are in each region. PostgreSQL holds instant answer and entity data. If you type in "duckduckgo", for example, you get something from Wikipedia and that comes from PostgreSQL. The PostgreSQL database has something like 100 sources of data. These are crawled by the backend and stored in the database.

Distributed caching system uses memcached. When something is cached it pushes out to all the other caching systems. There's no master using custom Perl solution. Caching is routed through Nginx so requests bypass the Perl backend completely if the data is in cache. There is no constrain by size, they can add as many cache machines as needed, the challenge is to figure out what to put in the cache so it will be useful and not to give bad results [20].

For web framework, Knockout is used in DDG which is a standalone JavaScript implementation of the Model-View-View-Model (MVVM) pattern with templates. The underlying principles are therefore a clear separation between domain data, view components and data to be displayed and the presence of a clearly defined layer of specialized code to manage the relationships between the view components [21].

Crawler: DuckDuckBot [4]

Revenue Generation:

DuckDuckGo generates revenue in two ways, while upholding privacy policy:

- Advertising
- Affiliate revenue

Advertising:

It is a myth that search engines need to track you to make money on Web search. When you type in a search, they can show an ad just based on that search term. For example, if you type in, "car" they show a car ad. That doesn't involve tracking because it is based on the keyword and not the person.

Advertising on DuckDuckGo takes the form of sponsored links that appear above search results. Sponsored links are currently syndicated through Yahoo!

Affiliate revenue:

DuckDuckGo is part of the affiliate programs of the ecommerce websites Amazon and eBay. When you visit those sites through DuckDuckGo and subsequently make a purchase, they receive a small commission.

This mechanism operates anonymously and there is no personally identifiable information exchanged between them and Amazon or eBay. These links are regular organic links (like any other link in our results) and these programs do not influence our ranking or relevancy functions in any way. That is, they are not advertising like paid placements or paid inclusions, and they only generate revenue from them if you ultimately find them relevant enough to end up purchasing an item.

3.6 Humkinar Search Engine:

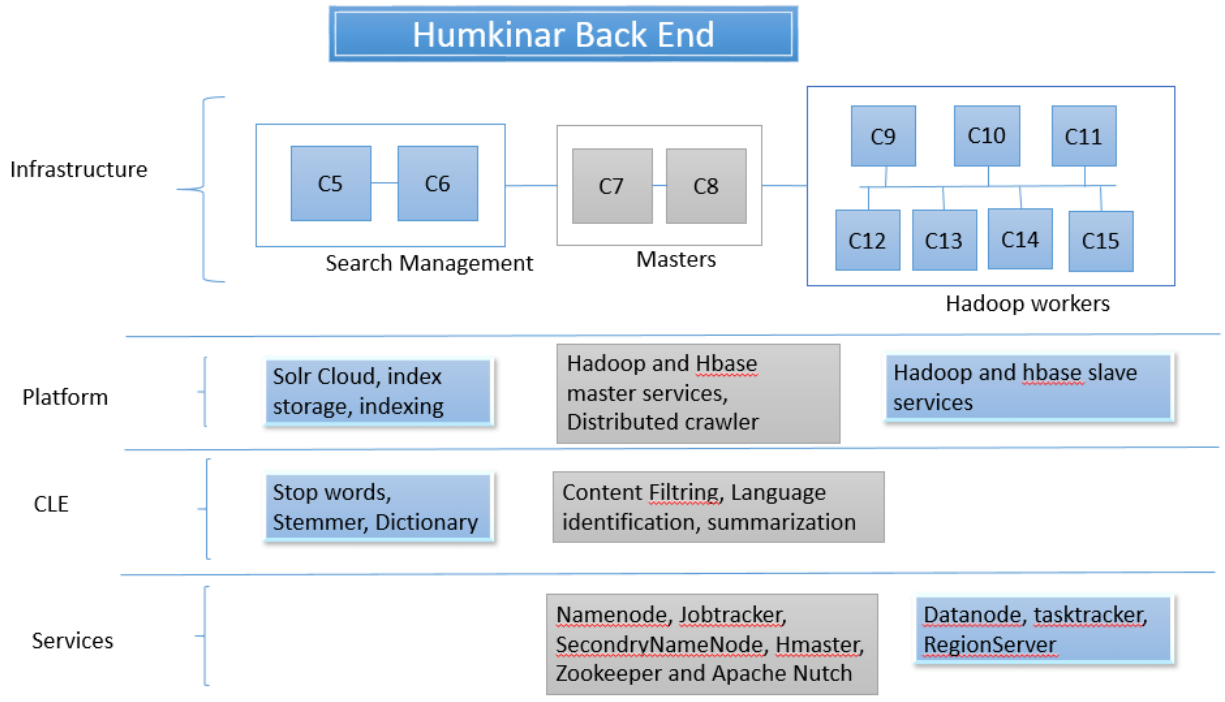

Figure 9: Humkinar Backend Architecture

The Humkinar cluster will consist of search management, masters and Hadoop worker's nodes. Apache Hbase, Solr, Zookeeper, Nutch and Hadoop services runs on these nodes. Frontend of Humkinar will also run on one of these systems.

In backend, the platform of Humkinar will consist of Apache Solr cloud for indexing and searching of the documents, Apache Hadoop in distributed mode for computation and storage, Apache Hbase a database inside Apache Hadoop which will store crawl data. Apache Nutch crawler in distributed mode which will run on Hadoop workers to take advantage of distributed computation while crawling the web. Apache ZooKeeper for centralized service, maintaining configuration information, naming, providing distributed synchronization and providing group services.

For Apache Solr security, we will use HAProxy which is used as a proxy server because by default, Solr listens on all interfaces. Due to this reason, it is very easy to update or even remove index of Solr with a single query. According to Solr reference manual, it is recommended that you should run Solr behind HAPROXY and allow Solr to listen on local interface (localhost) only. But as our crawler jobs will be running on different systems, so every node will access Solr so that documents can update in Solr.

For frontend, we use PHP Codeigniter framework, jQuery, Javascript, Bootstrap framework for responsiveness of screen sizes and redis (in-memory data structure store) to get data instantly.

Humkinar will also provide Language Identification, Content Filtering, Text Summarization, Search Suggestions, On-screen keyboard, Book Search and etc.

Humkinar architecture is mostly similar to other search engines like Yahoo, Baidu because Yahoo also use Hadoop for computation and storage, Hbase as a database, Zookeeper for group services etc. We are not using Software Defined Network (SDN).

4 Summary:

	Humkinar	**Yahoo**	**Google**	**Common Search**	**Duck Duck Go**	**Baidu**
Distributed Computation & Storage	Apache Hadoop	Apache Hadoop	GFS/ Colossus	Apache Spark	Amazon EC2	Apache Hadoop
Database	Apache Hbase	Apache Hbase	Big Table	No	PostgreSQL	Hbase
Indexer	Apache Solr	N/A	Caffeine	Elastic Search	Apache Solr	N/A
Crawler	Apache Nutch	Yahoo Slurp and Bing Bot	Google Bot	Common Crawl	DuckDuckBot	Baiduspider
Page	Available	Availab	Google	Static Ranking	Available	Availab

Rank		le	Page Rank			le
Coordination	Apache Zookeeper	Apache Zookeeper	Chubby & Borg & Omega	AWS Cloud Formation	N/A	N/A
Frontend Framework	PHP CodeIgniter	YUI	N/A	markdown & pelican	YUI & Knockout	N/A
In-memory data structure store	Redis	N/A	N/A	N/A	MemCached	N/A
Software Define Network	No	N/A	B4	No	N/A	N/A

5 Comparison with Language-focused Search Engine

In this section, we look at some popular Persian and Arabic search engines. In order to compare our proposed search engine with similar ones on the basis of language, we put together a list of features that ours offers and test other search engines for their presence. These features are listed below.

1. **Language Identification:** Does the search engine use language identification to focus the search on a specific subset of the Internet and/or rank search results different depending upon the predominant language in them?
2. **Content Filtering:** Does the search engine adopt any content filtering techniques in order to screen objectionable written content?
3. **Text Summarization:** Does the search engine offer text summarization to the user to allow quick skimming of search results' contents without having to visit them?
4. **Search Suggestions:** Does the search engine present suggestions for similar search terms to the user as he is entering his query?
5. **On-screen keyboard:** Does the search engine offer the user an on-screen keyboard in order to type queries containing words written in the Arabic script?
6. **Book Search:** Does the search engine allow the user to search within books and literary material for academic purposes?
7. **Content-rich Homepage:** Does the search engine present current news and updates on its homepage in order to keep its users well-informed?
8. **Autonomy:** Is the search engine independent or does it rely on another search engine for processing queries?

5.1 Persian Search Engines

In this section, we look at some popular Iranian search engines developed primarily for catering to users who are either residents of Iran or wish to search the internet using the Persian language.

5.1.1 Yooz

Yooz [22] is one of the well-known search engines in Iran which was founded in 2010. This Iranian search engine made some noise when it was announced that the company had received 2 million dollars of funding with the help of the government. According to the CEO of Yooz, 25 million Iranians use Google's search engine daily and this is higher than countries such as Germany and France. Recently, Yooz announced that they have stored the data of more than one billion Iranian websites in their servers and they are planning to have 30 percent share of the searches done by Iranians. Aside from being able to search for specific keywords, users on Yooz, can also search for news, blogs and images. Iran's Ministry of Communication and IT has said the search engine is capable of supporting up to one billion Persian websites, and is claimed to have currently indexed over 1 billion web pages. Yooz is Persian for cheetah. Yooz search engine has 100,000 hits and more than 60,000 searchers per day and is Iran's 4th most visited website after Google, Parsijoo and Bing. Alexa ranks Yooz 75,820 worldwide (down 21,205).

Yooz does offer language identification in order to rank relevant pages higher. It also provides search suggestions as you type your query, and has a content-rich homepage which updates you on the latest news and happenings. Since it does not rely on Google for searching, it is an autonomous, self-sustaining search engine. However, it does not support content-filtering, text summarization, an on-screen keyboard, or book search.

A screenshot of its user interface is shown in the figure below.

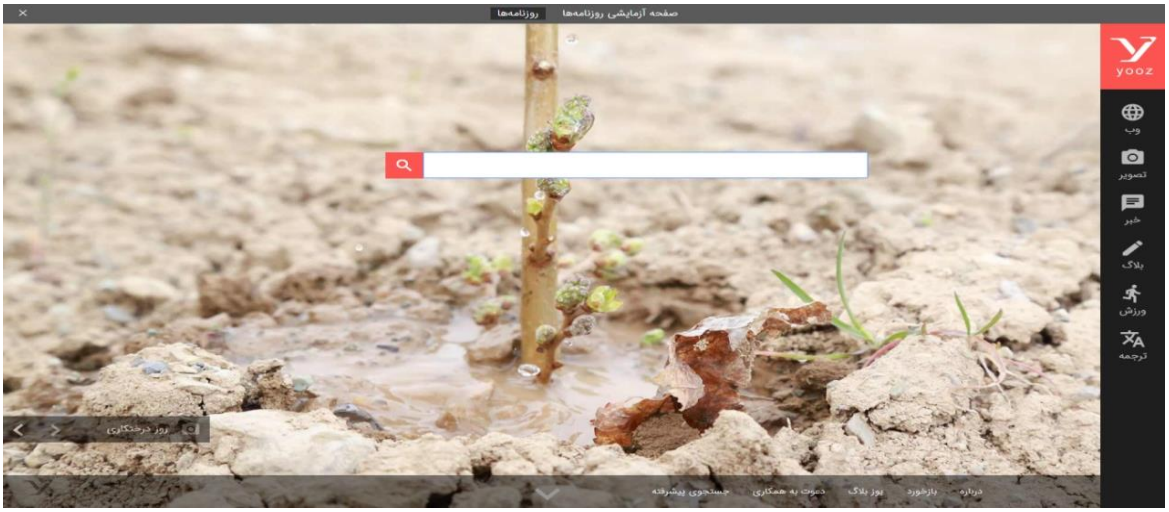

Figure 10: Yooz Interface

5.1.2 Rismoon

Rismoon [23] translated into English as “string”, claims to be the first search engine for Persian content. The website was launched in 2004 by a private company which claims it has not received any funding by other private or governmental organizations. Rismoon has a feature called “Zoom Search” which lets the user to search for the keywords in a specific topic or area. The website also stores the data of Iran’s telephone directory of businesses known to most locals as “118”. Alexa ranks Rismoon 527,574 worldwide (up 334,671).

Rismoon does offer language identification, an on-screen keyboard. Since it does not rely on Google for searching, it is an autonomous, self-sustaining search engine. However, it does not support content-filtering, text summarization, search suggestions, book search, or a content-rich homepage.

A screenshot of its user interface is shown in the figure below.

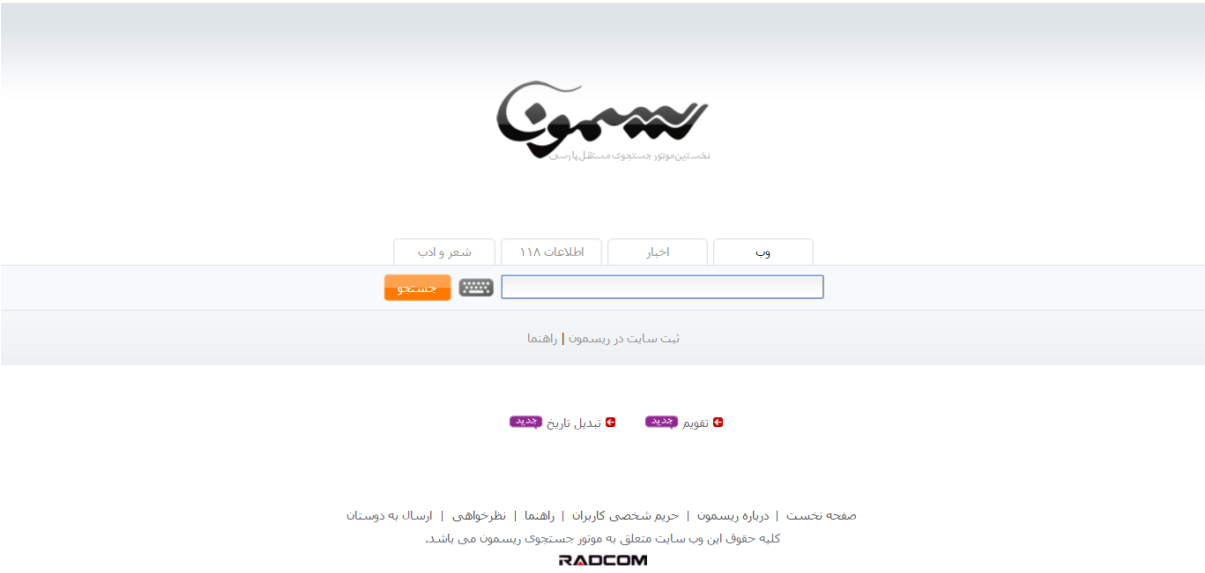

Figure 11: Rismoon Interface

5.1.3 Salam

Salam [24] is a kind of Iranian “Meta Search Engine”. Salam searches for the keywords in a couple of search engines simultaneously and shows an optimal combination of all of them. Salam benefits from an Artificial Intelligence system and uses advanced mathematical models to bring more accurate results for Persian keywords. Salam is backed by Bayan, a software company that also offers other domestic services such as blog, email and file hosting. Bayan is currently developing a Persian search engine called Zal. Alexa ranks Salam 111,581 worldwide (up 4,049).

Salam does offer language identification, content-filtering, and search suggestions. Since it does not rely on any one particular search engine for searching, it can be said to be an autonomous search engine to some degree. However, it does not support text summarization, an on-screen keyboard, book search, or a content-rich homepage.

A screenshot of its user interface is shown in the figure below.

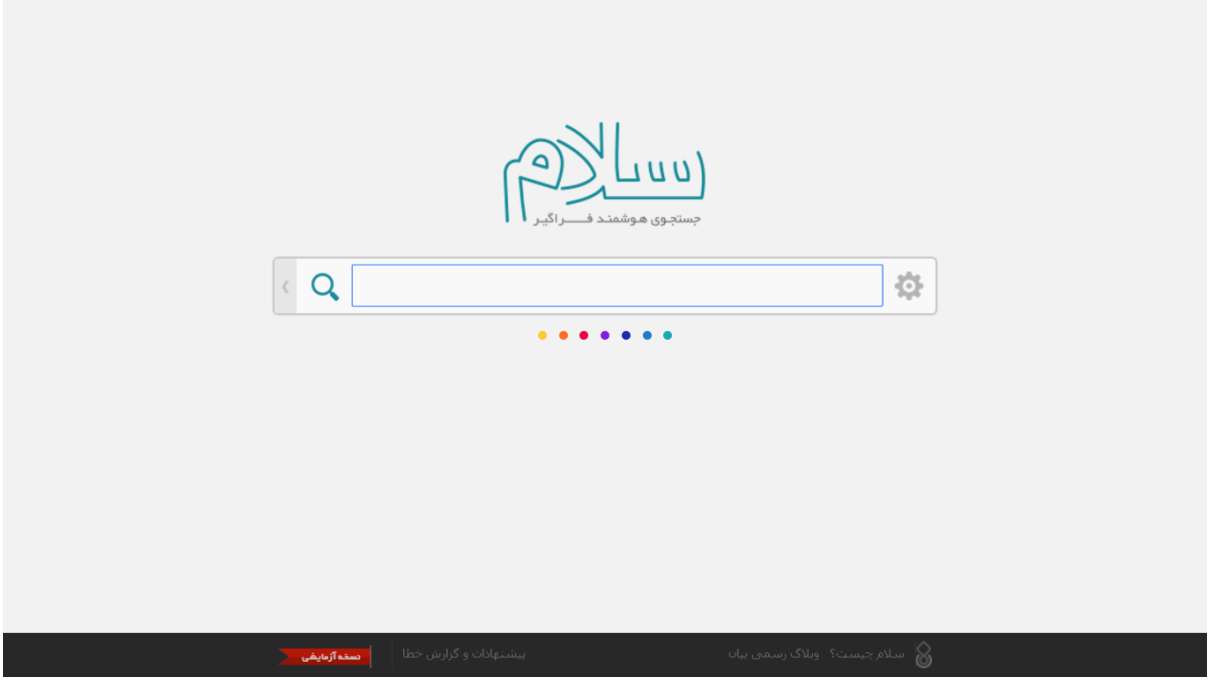

Figure 12: Salam Interface

5.1.4 Parsijoo

Parsijoo [25] is yet another Persian search engine developed by Iranians. The beta version of the website was available in 2010. Parsijoo became commercially available in November 2015 with over one billion indexed web pages. The website has the ability to search for various content such as text, audio, video, image and downloadable files and apps. Some of the other features include an online dictionary, map, job board and news section. Users can also search for a specific item on the Bazaar section of Parsijoo and see the results in 20 Iranian e-commerce websites. Alexa ranks Parsijoo 16,759 worldwide (up 2,352).

Parsijoo does offer language identification, search suggestions, book search, and a content-rich homepage. Since it does not rely on Google for searching, it is an autonomous, self-sustaining search engine. However, it does not support content-filtering, text summarization, or an on-screen keyboard.

A screenshot of its user interface is shown in the figure below.

Figure 13: Parsijo Interface

5.2 Arabic Search Engines

In this section, we look at some popular Arabic search engines developed primarily for catering to users who are either residents of Saudi Arabia or wish to search the internet using the Arabic language.

5.2.1 Yamli

Yamli [26] is an Internet start-up focused on addressing the problems specific to the Arabic web. Yamli currently offers two main products: the smart Arabic keyboard, and Yamli Arabic Search. The smart Arabic keyboard allows users to type Arabic without an Arabic keyboard from within their web browser. This technology is based on a real-time transliteration engine which converts words typed with Latin characters to their closest Arabic equivalent. Yamli Arabic search is a search engine focused on providing more relevant search results for an Arabic query by expanding it to its most frequently used Latin representations.

Yamli's typing technology originated during the July 2006 war in Lebanon, when co-founder Habib Haddad was looking for news about the war online. Without access to an Arabic keyboard and not being used to one, Haddad had difficulties finding up-to-date information, which is generally first available in Arabic. After working on prototypes of an Arabic transliteration engine for several months, Haddad, along with co-founder Imad Jureidini, founded Language Analytics LLC in July 2007. In November 2007, Language Analytics launched Yamli.com, featuring the Smart Arabic Keyboard to allow users to type Arabic without an Arabic keyboard, as well as a basic search engine front end to Google. This first search engine did not include the expanded query capabilities of Yamli Arabic Search. The typing technology was made to third-party websites in March 2008 in the form of a free API. The search engine was updated to include the query expansion capabilities in December 2008. Alexa ranks Yamli 14,830 worldwide (down 1,442).

Yamli does not offer language identification, content-filtering, text summarization, search suggestions, book search, or even a content-rich homepage. Since it is based on Google, it is not autonomous.

A screenshot of its user interface is shown in the figure below.

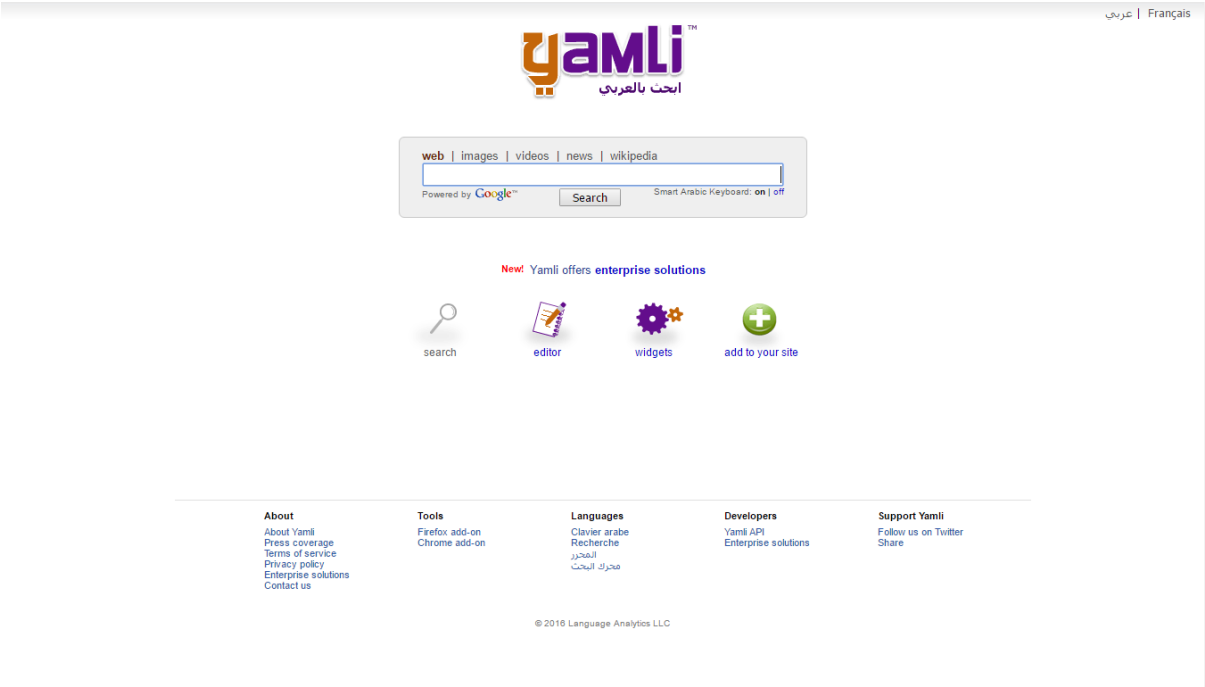

Figure 14: Yamli Interface

5.2.2 Yoolki

Yoolki [27] is another Arabic search engine. Alexa ranks Yoolki 6,550,356 worldwide (down 1,518,639).

Yoolki does not offer language identification, content-filtering, text summarization, search suggestions, an on-screen keyboard, book search, or even a content-rich homepage. Since it is based on Google, it is not autonomous.

A screenshot of its user interface is shown in the figure below

Figure 15: Yoolki Interface

5.2.3 Eiktub

Eiktub [28] is another Arabic search engine. Alexa ranks Eiktub 3,853,686 worldwide (up 396,060).

Eiktub does not offer language identification, content-filtering, text summarization, search suggestions, an on-screen keyboard, book search, or even a content-rich homepage. Since it is based on Google, it is not autonomous.

A screenshot of its user interface is shown in the figure below.

Figure 16: Eiktub Interface

6 Summary

A summary of the comparison presented in the above sections is depicted in the table below. Our proposed search engine (Humkinar) is highlighted red.

	Humkinar	***Yooz***	***Rismoon***	***Salam***	***Parsijoo***	***Yamli***	***Yoolki***	***Eiktub***
Language ID	✔	✔	✔	✔	✔			
Content Filtering	✔			✔				
Text Summariz-ation	✔							
Search Suggestions	✔	✔		✔	✔			
On-screen Keyboard	✔		✔					
Book Search	✔				✔			
Content-rich Homepage	✔	✔			✔			
Autonomy	✔	✔	✔	✔	✔			

7 Reference:

[1] http://www.wordstream.com/articles/internet-search-engines-history

[2] https://en.wikipedia.org/wiki/Yahoo!_Search

[3] http://yahoohadoop.tumblr.com/

[4] https://www.keycdn.com/blog/web-crawlers/

[5] https://en.wikipedia.org/wiki/YUI_Library

[6] http://www.irkawebpromotions.com/search-engines/yahoo/

[7] https://en.wikipedia.org/wiki/Google_Search

[8] https://www.wired.com/2012/07/google-colossus/

[9] http://www.slideshare.net/AditiTechnologies/google-architecture-breaking-it-open

[10] https://googleblog.blogspot.com/2010/06/our-new-search-index-caffeine.html

[11] http://thenewstack.io/googles-data-architecture-and-what-it-takes-to-work-at-scale/

[12]https://docs.google.com/presentation/d/1OvJStE8aohGeI3y5BcYX8bBHwoHYCPu99A3KTTZElr0/edit#slide=id.gb74341dde_0_64

[13]http://www.networkcomputing.com/networking/inside-googles-software-defined-network/512240144

[14] https://en.wikipedia.org/wiki/PageRank

[15] https://www.searchenginejournal.com/seo-guide/panda-penguin-hummingbird/

[16] https://about.commonsearch.org/developer/architecture

[17]https://conferences.oreilly.com/strata/big-data-conference-sg-2015/public/schedule/detail/44755

[18] http://www.alluxio.org/docs/master/en/

[19] https://en.wikipedia.org/wiki/DuckDuckGo

[20] http://highscalability.com/blog/2013/1/28/duckduckgo-architecture-1-million-deep-searches-a-day-and-gr.html

[21] https://en.wikipedia.org/wiki/Knockout_(web_framework)

[22] www.yooz.ir

[23] www.rismoon.ir

[24] www.salam.ir

[25] www.parsijoo.ir

[26] www.yamli.com

[27] www.yoolki.com

[28] www.eiktub.com

[29] https://en.wikipedia.org/wiki/AdSense

[30] http://www.linksandlaw.com/technicalbackground-pagerank.htm

[31] https://montfort.io/google-rank-brain-what-is-it-and-how-will-it-affect-seo/

[32] http://searchengineland.com/library/google/google-pigeon-update

[33] http://searchengineland.com/everything-need-know-pigeon-algorithm-211771